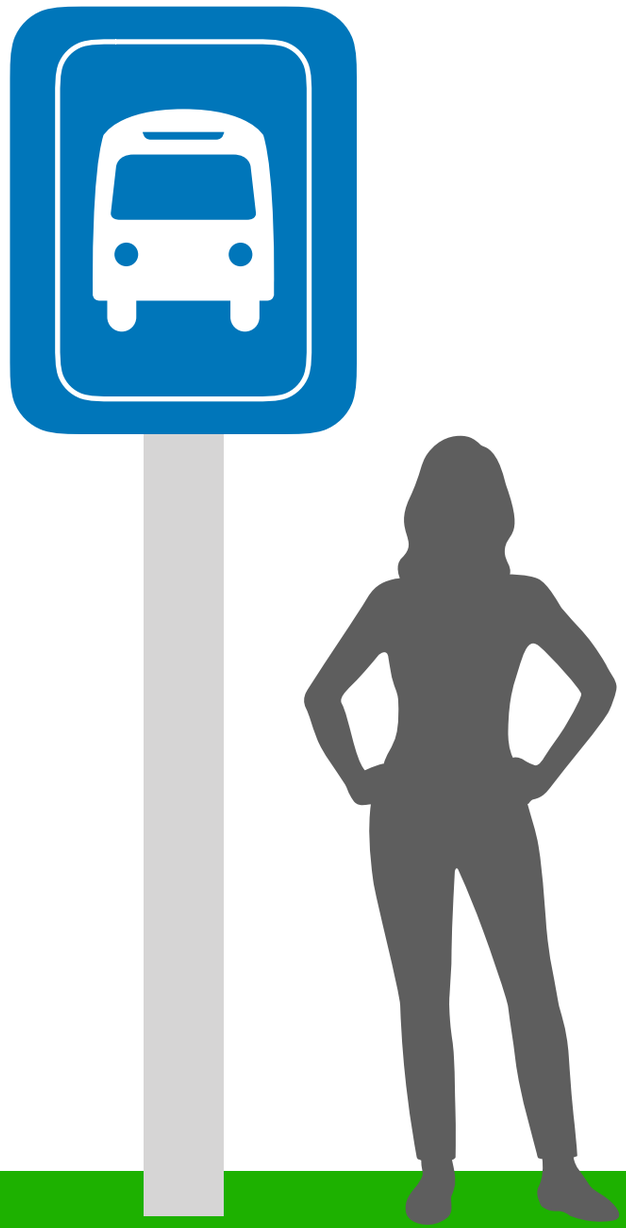
Lecture 12

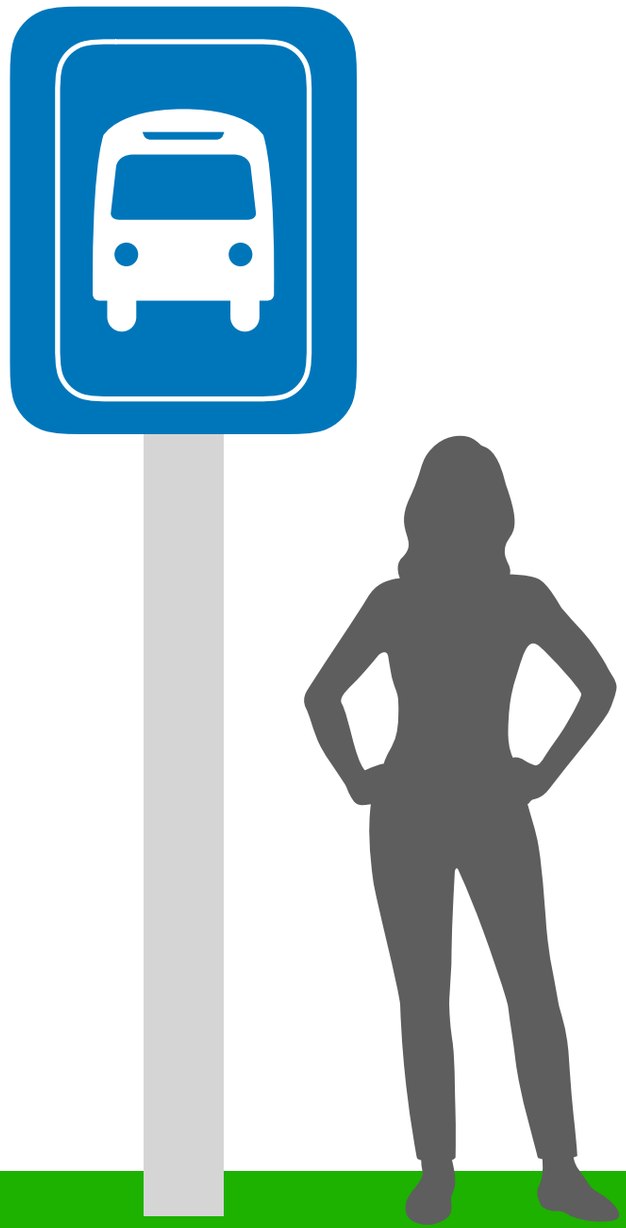
Bayesian inference
Botany 563 — Spring 2022

- **Previous class check-up:**
    - We reviewed the likelihood methods and their strengths and weaknesses
- **Learning Objectives:** At the end of today's session, you will be able to
    - Explain the main characteristics of Bayesian inference for phylogenetics
    - Understand the role that priors, sample size and convergence play in the performance of Bayesian inference
- **Pre-class work**
    - Read HAL 1.4 and quiz

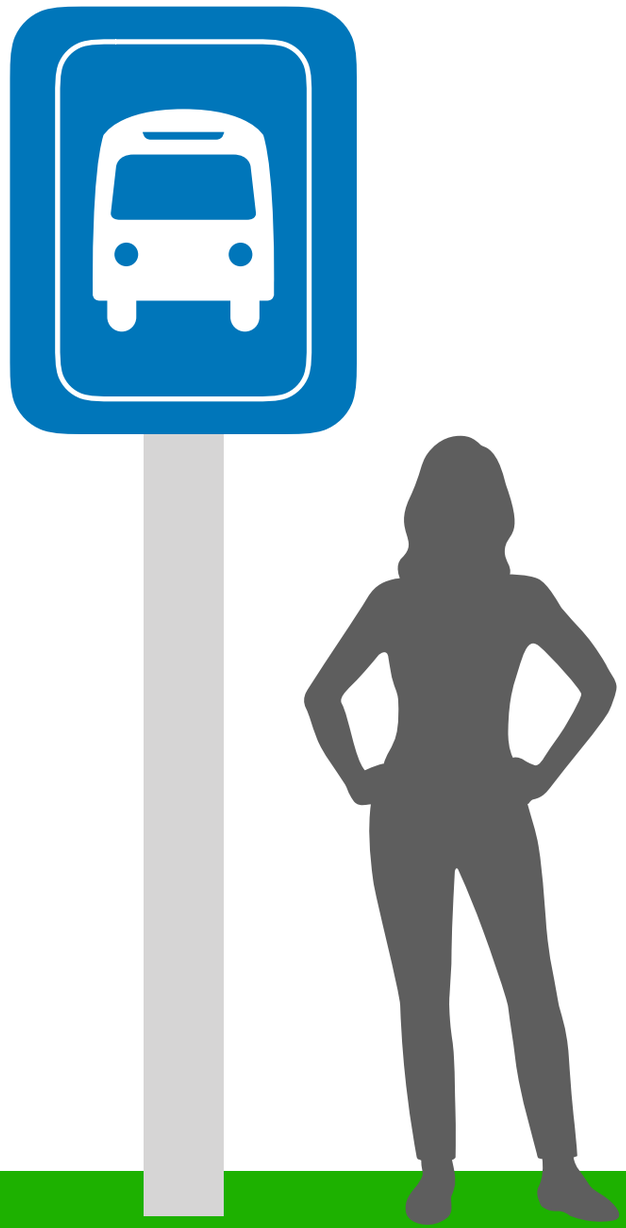
A familiar example



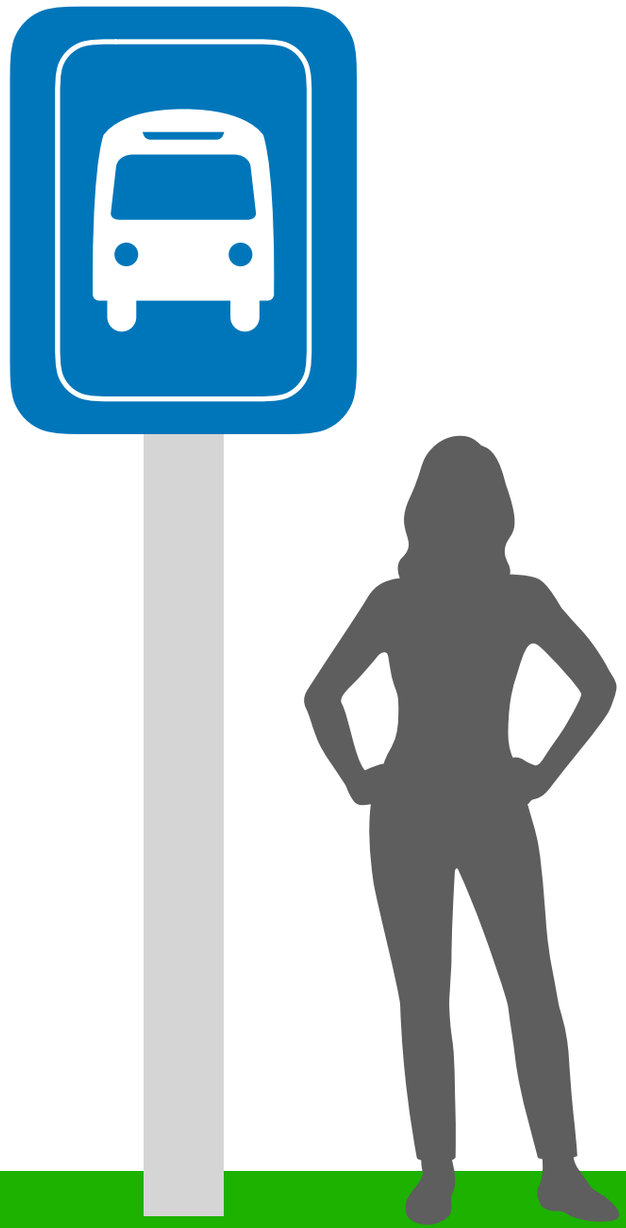
A familiar example



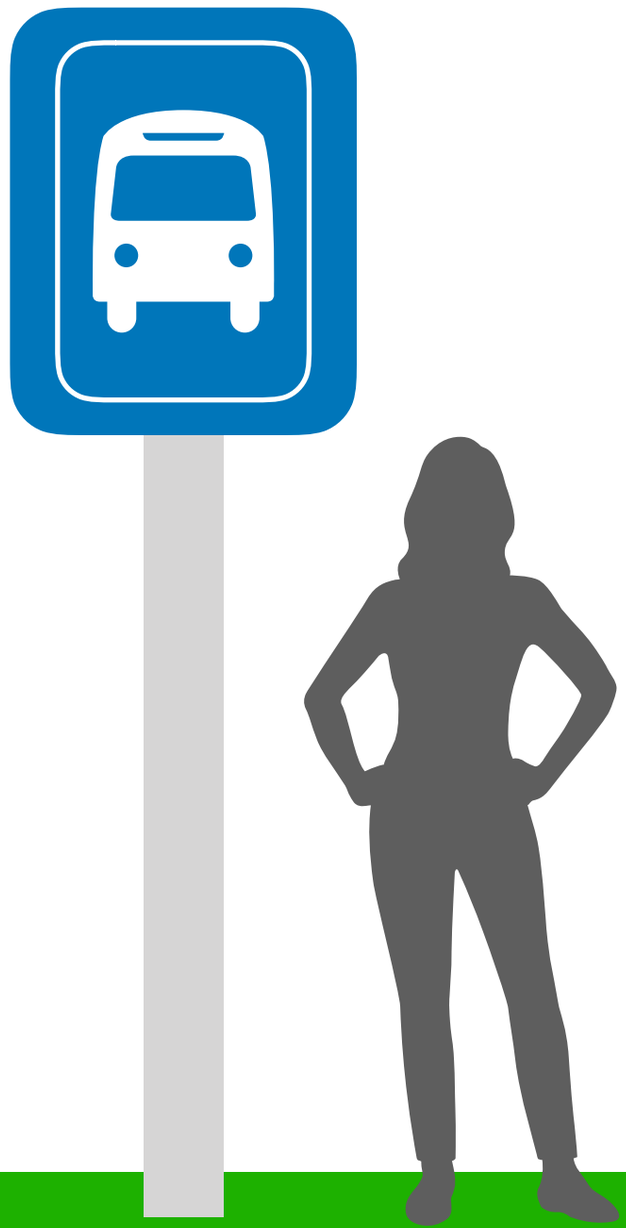
A familiar example



A familiar example



A familiar example



$$X_i = \text{Number of bikes seen on day } i$$

$$X_i \sim Poisson(\lambda)$$

$$P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

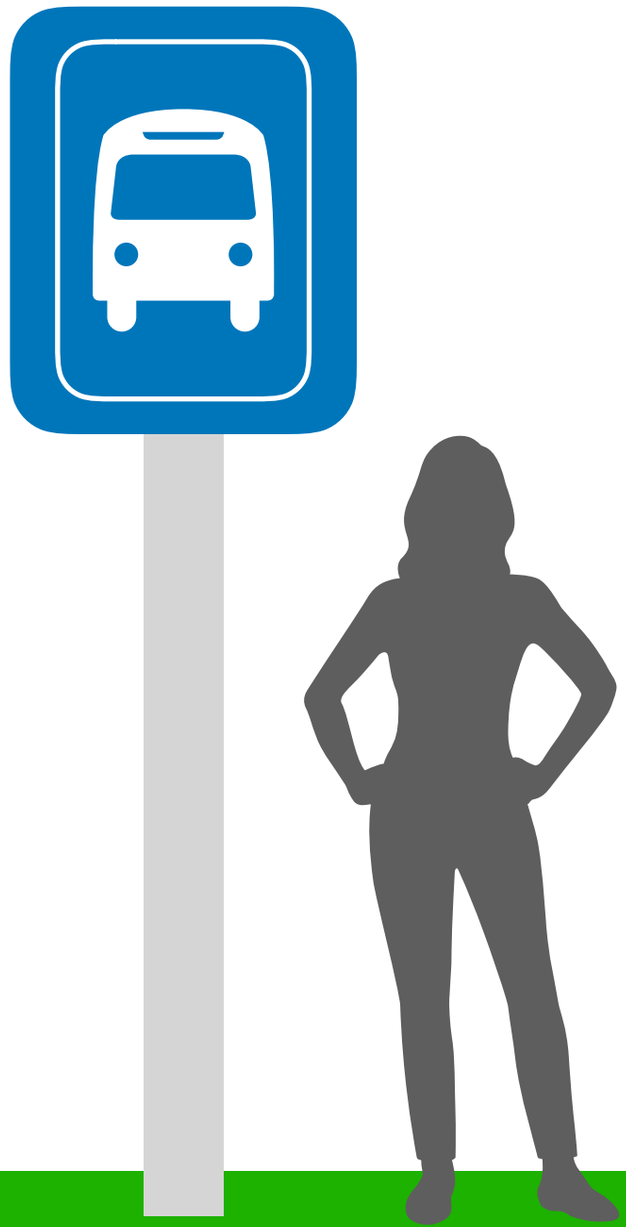
We collect data on n days:

$$X_1, X_2, \ldots, X_n$$

We want the MLE for λ

1. We need the likelihood function
2. We need to maximize the likelihood function

A familiar example



$$X_i = \text{Number of bikes seen on day } i$$

$$X_i \sim Poisson(\lambda)$$

$$P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We collect data on n days:

$$X_1, X_2, \ldots, X_n$$

Sequences

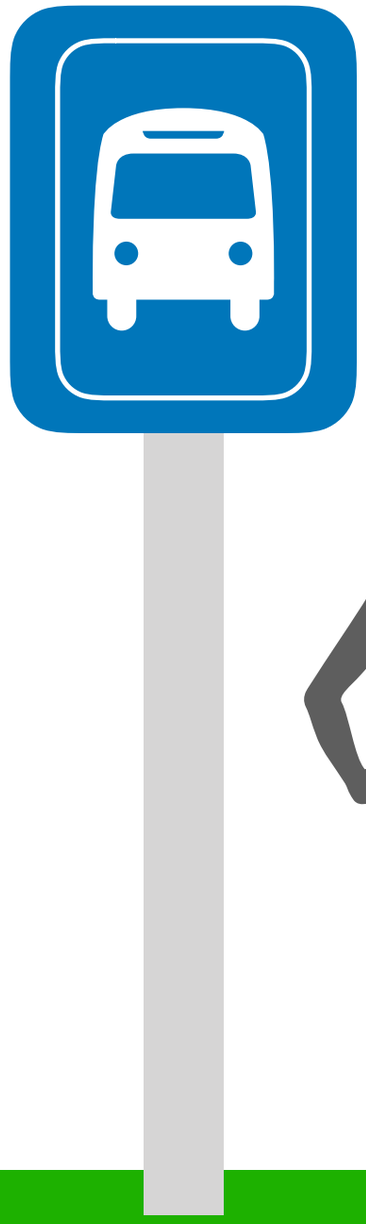
We want the MLE for λ

Tree, BL, Q

1. We need the likelihood function
2. We need to maximize the likelihood function

A familiar example

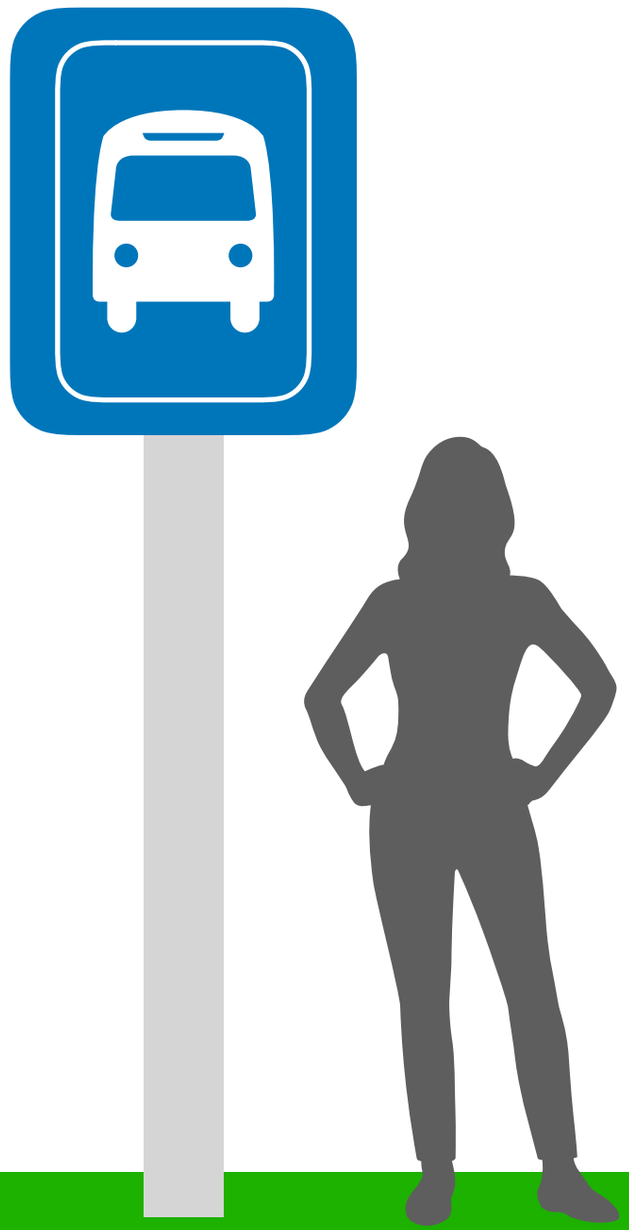
$$X_i = \text{Number of bikes seen on day } i$$



$$X_i \sim Poisson(\lambda) \;\; P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We collect data on n days: $X_1, X_2, \ldots, X_n$

A familiar example



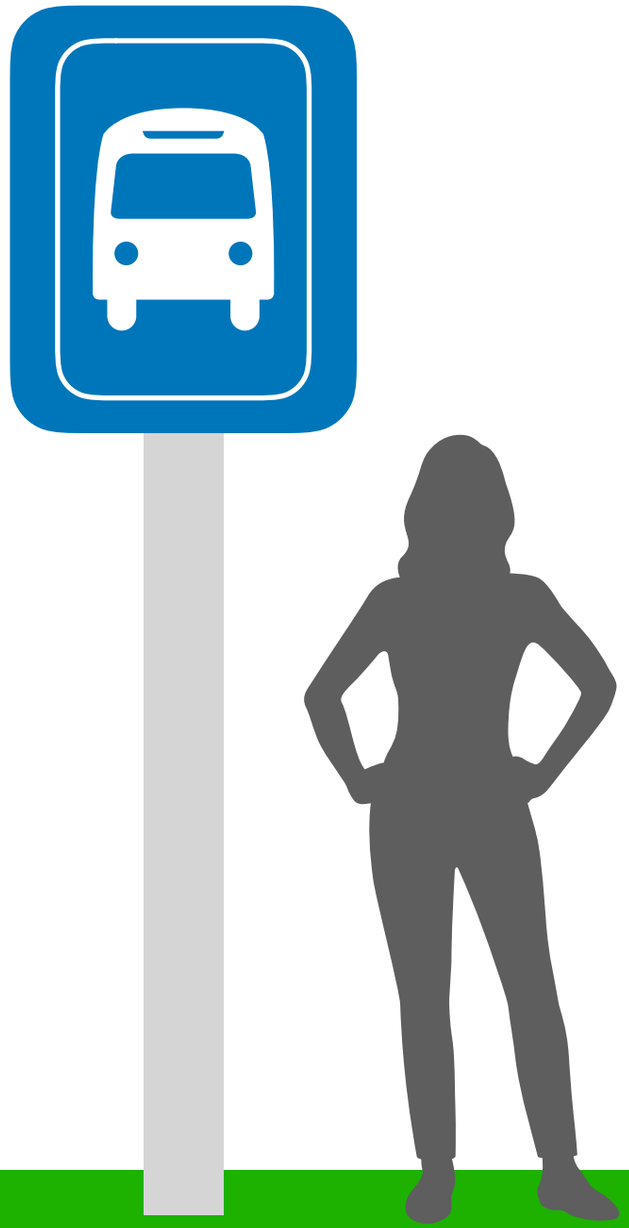
$$X_i = \text{Number of bikes seen on day } i$$

$$X_i \sim Poisson(\lambda) \;\; P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We collect data on n days: $X_1, X_2, \ldots, X_n$

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

A familiar example



$$X_i = \text{Number of bikes seen on day } i$$

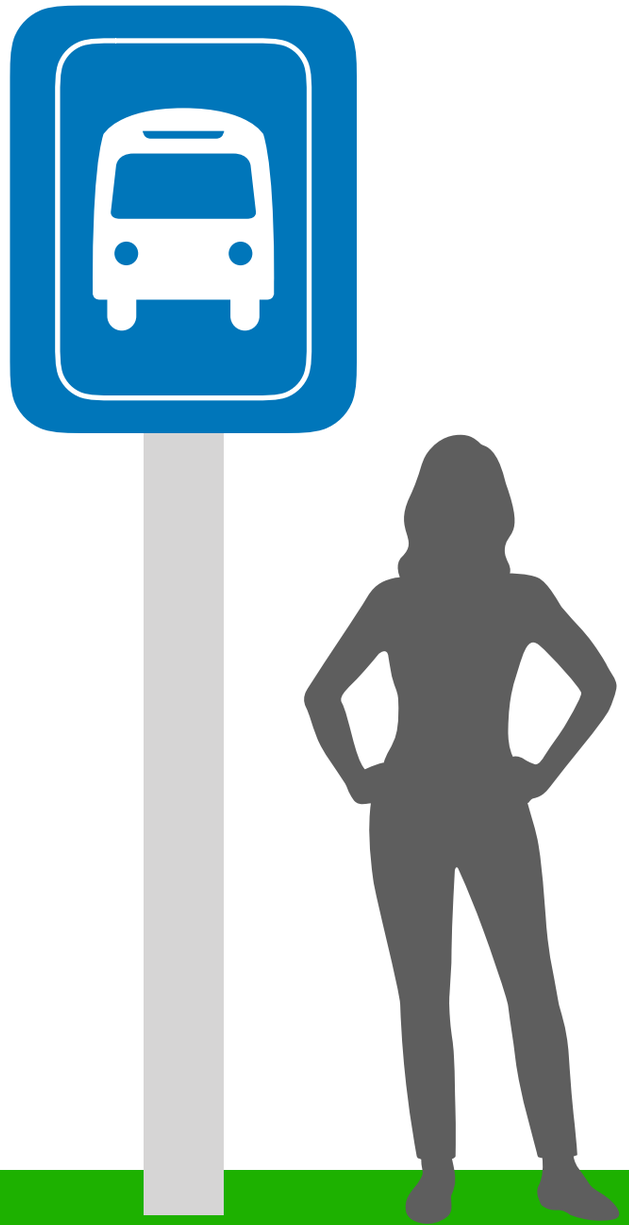
$$X_i \sim Poisson(\lambda) \;\; P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We collect data on n days: $X_1, X_2, \ldots, X_n$

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

$$l(\lambda) = -n\lambda + (\sum x_i)\log(\lambda) - \log(\prod x_i!)$$

A familiar example



$$X_i = \text{Number of bikes seen on day } i$$

$$X_i \sim Poisson(\lambda) \;\; P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

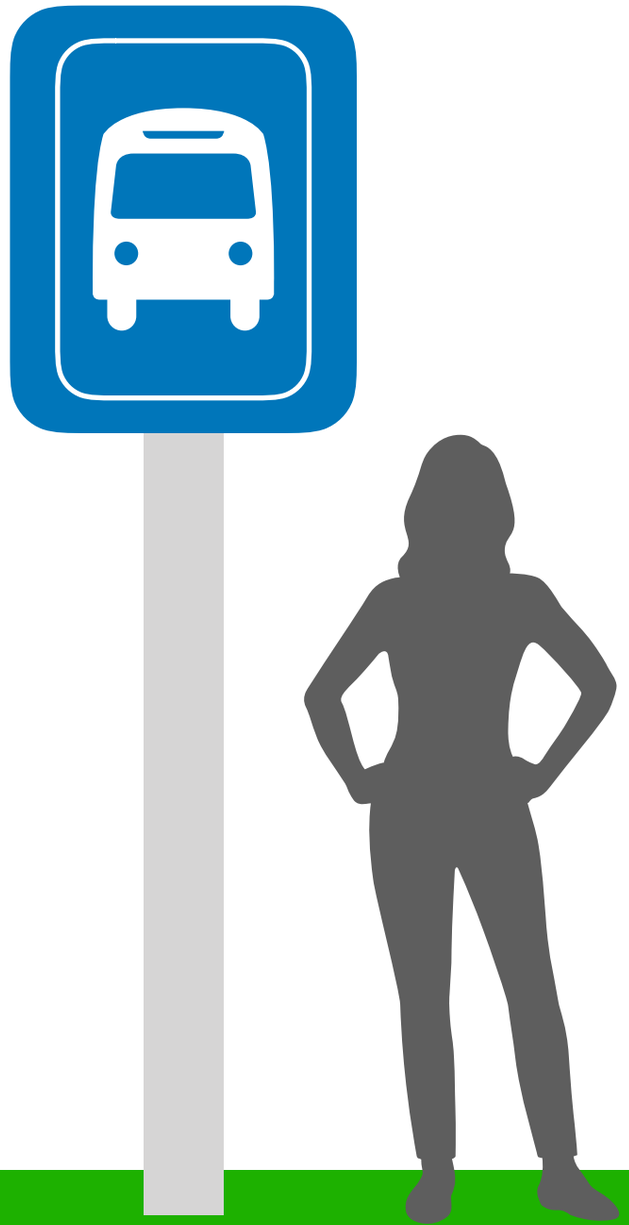
We collect data on n days: $X_1, X_2, \ldots, X_n$

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

$$l(\lambda) = -n\lambda + (\sum x_i)\log(\lambda) - \log(\prod x_i!)$$

$$l'(\lambda) = -n + \frac{\sum x_i}{\lambda} = 0$$

A familiar example



$$X_i = \text{Number of bikes seen on day } i$$

$$X_i \sim Poisson(\lambda) \;\; P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We collect data on n days: $X_1, X_2, \ldots, X_n$

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

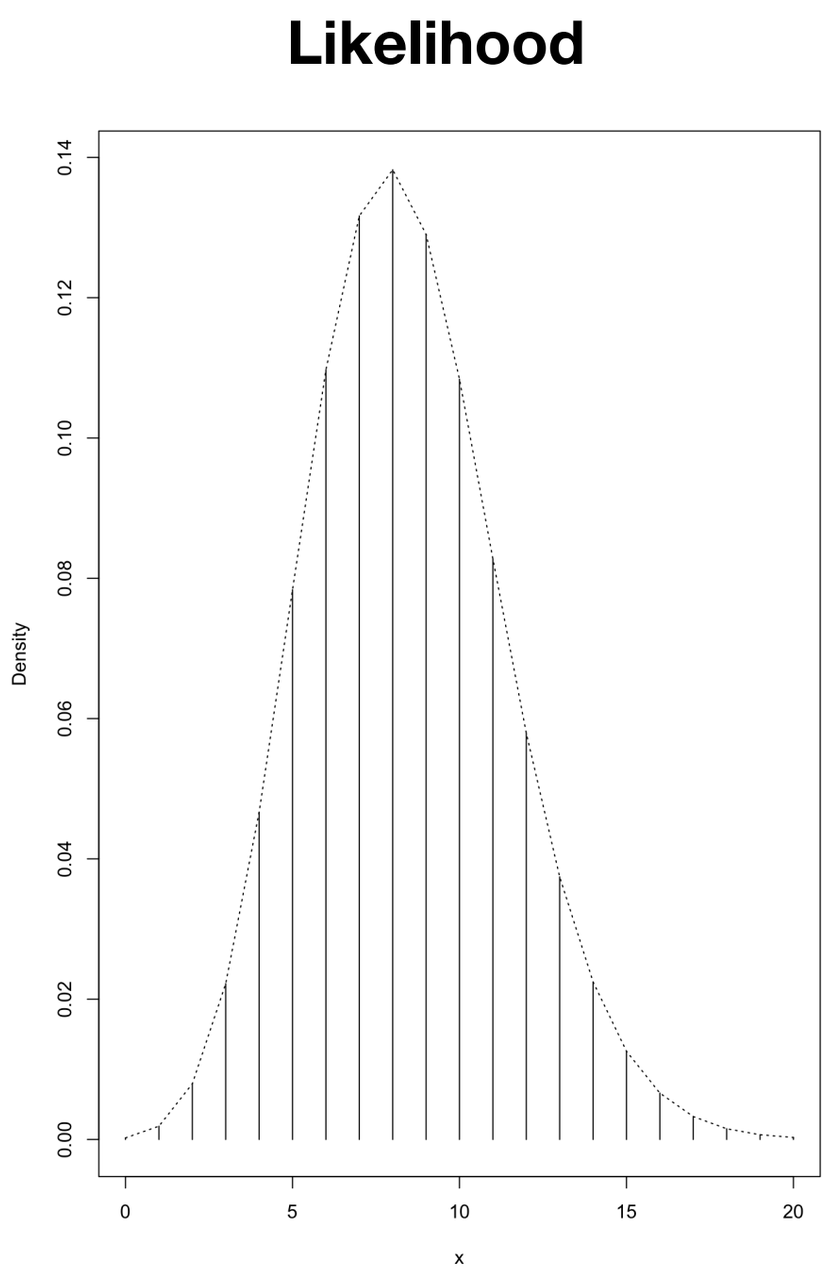
$$l(\lambda) = -n\lambda + (\sum x_i)\log(\lambda) - \log(\prod x_i!)$$

$$l'(\lambda) = -n + \frac{\sum x_i}{\lambda} = 0 \rightarrow \hat{\lambda} = \frac{\sum x_i}{n}$$

MLE

Bayesian: Likelihood 2.0

Information in the data

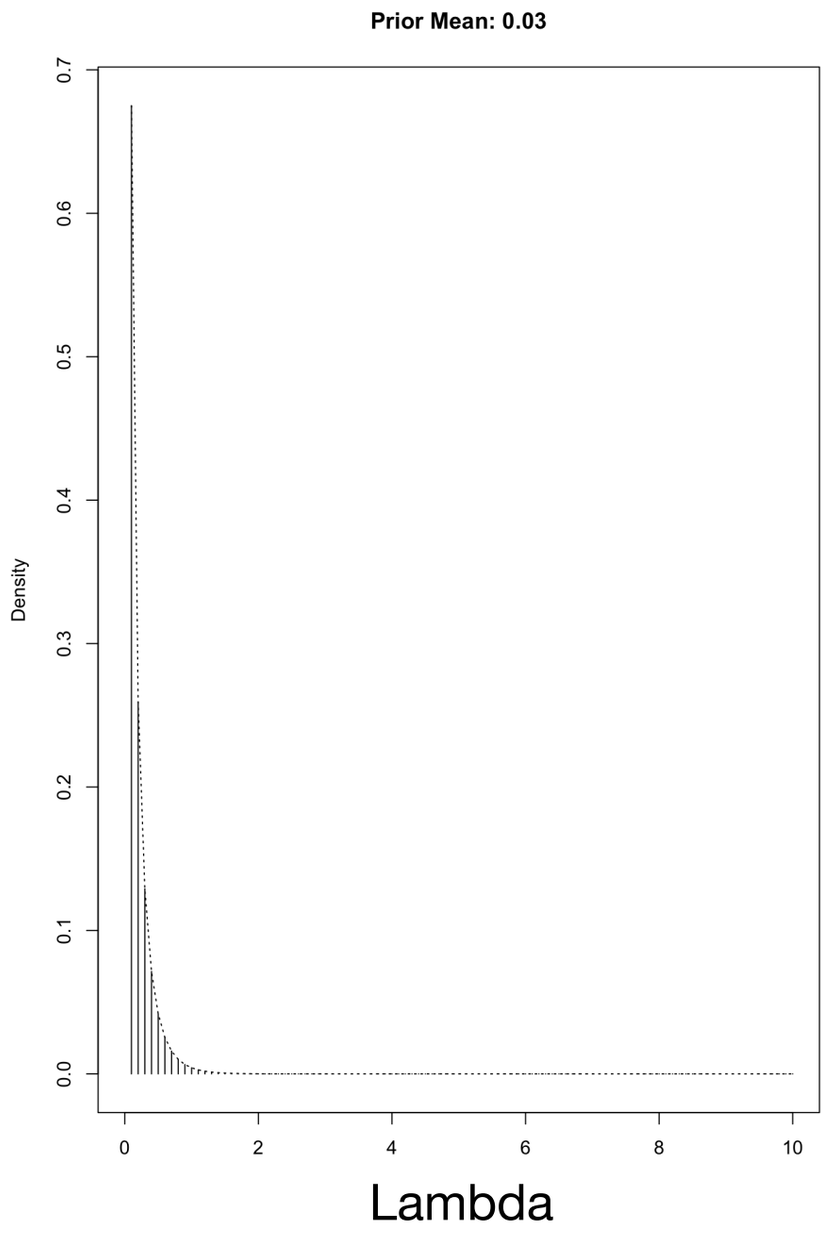


$$\hat{\lambda} = 8.4$$

Bayesian: Likelihood 2.0

Your knowledge

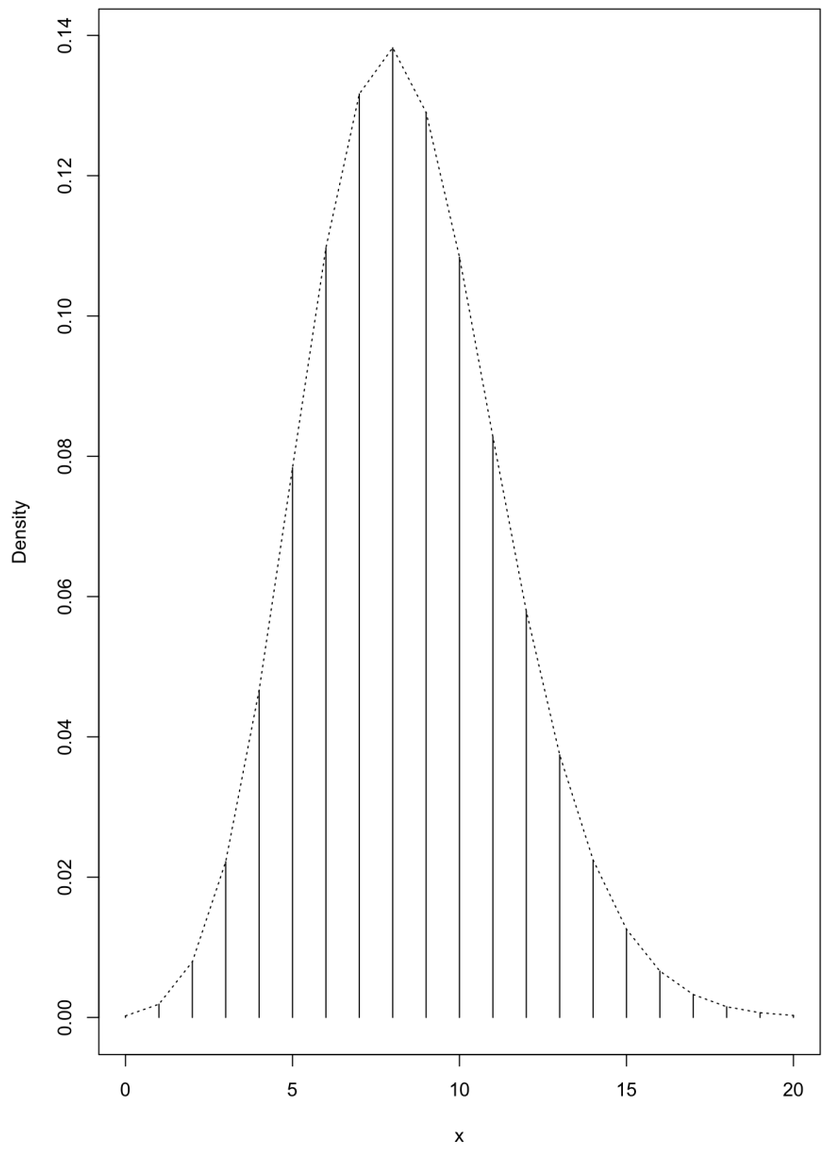
Prior



$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

Information in the data

Likelihood

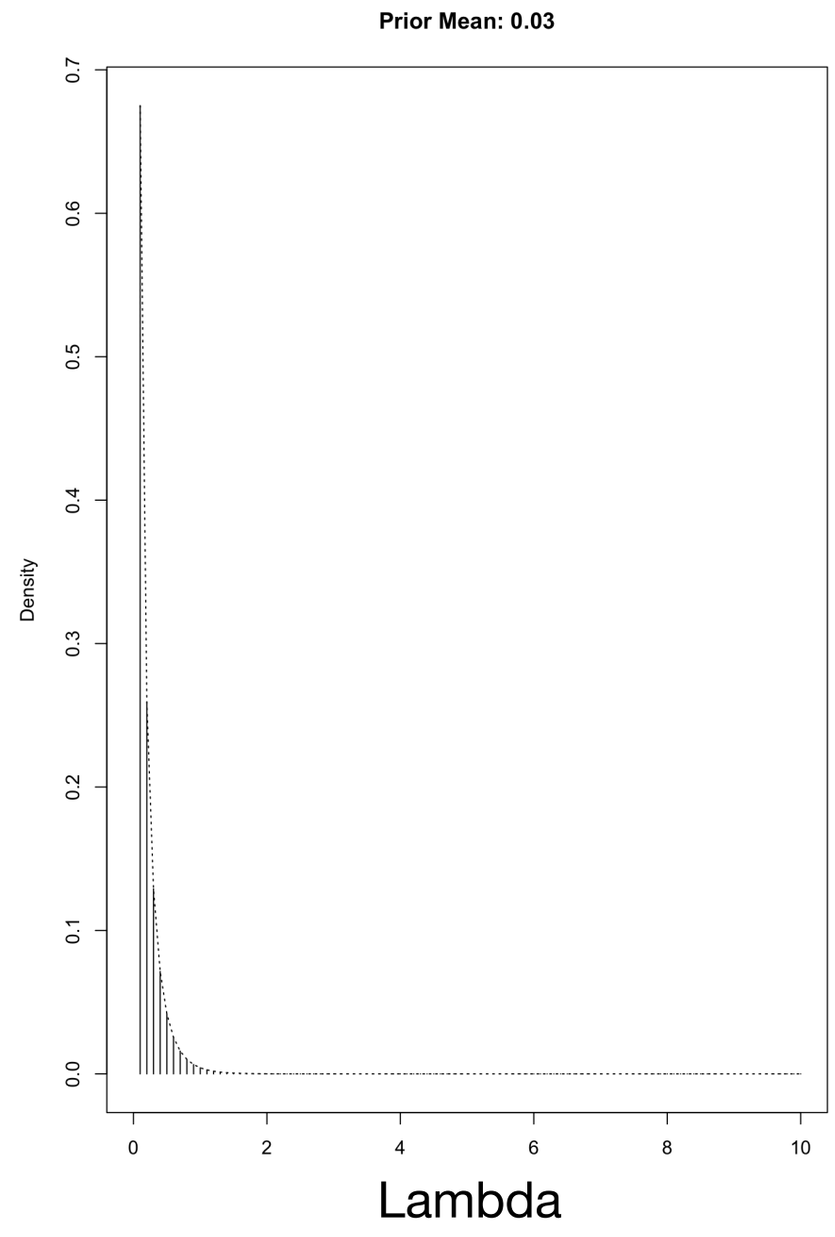


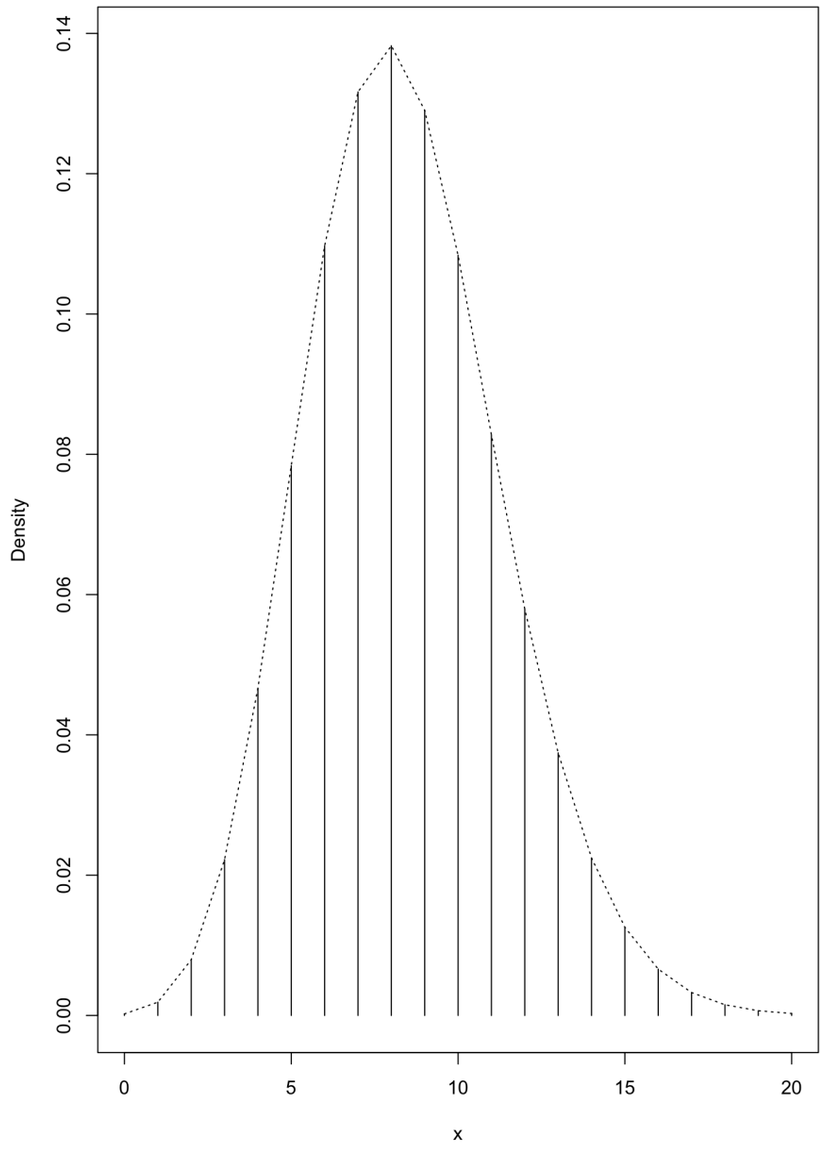
$$\hat{\lambda} = 8.4$$

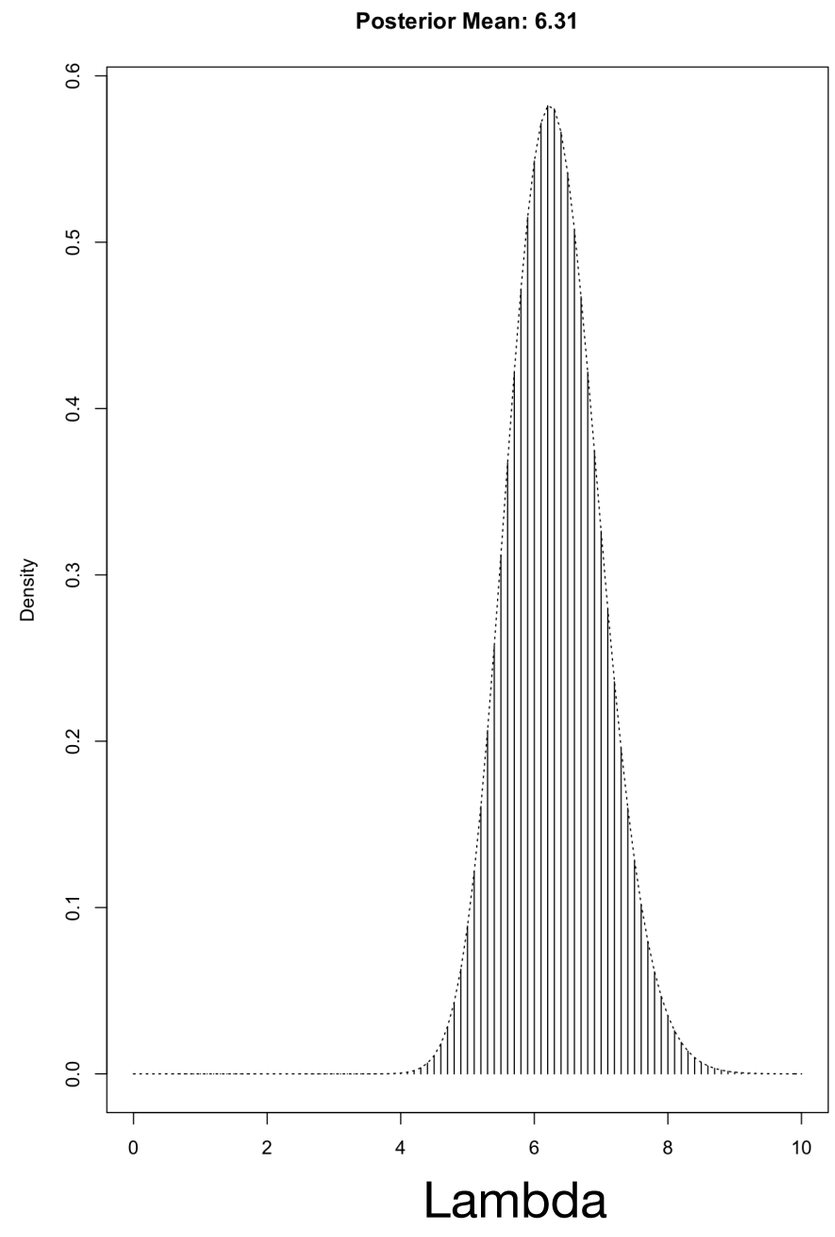
Bayesian: Likelihood 2.0

Your knowledge | Information in the data | Inference on parameter

Prior + **Likelihood** = **Posterior**







$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$

$\hat{\lambda} = 8.4$

$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$

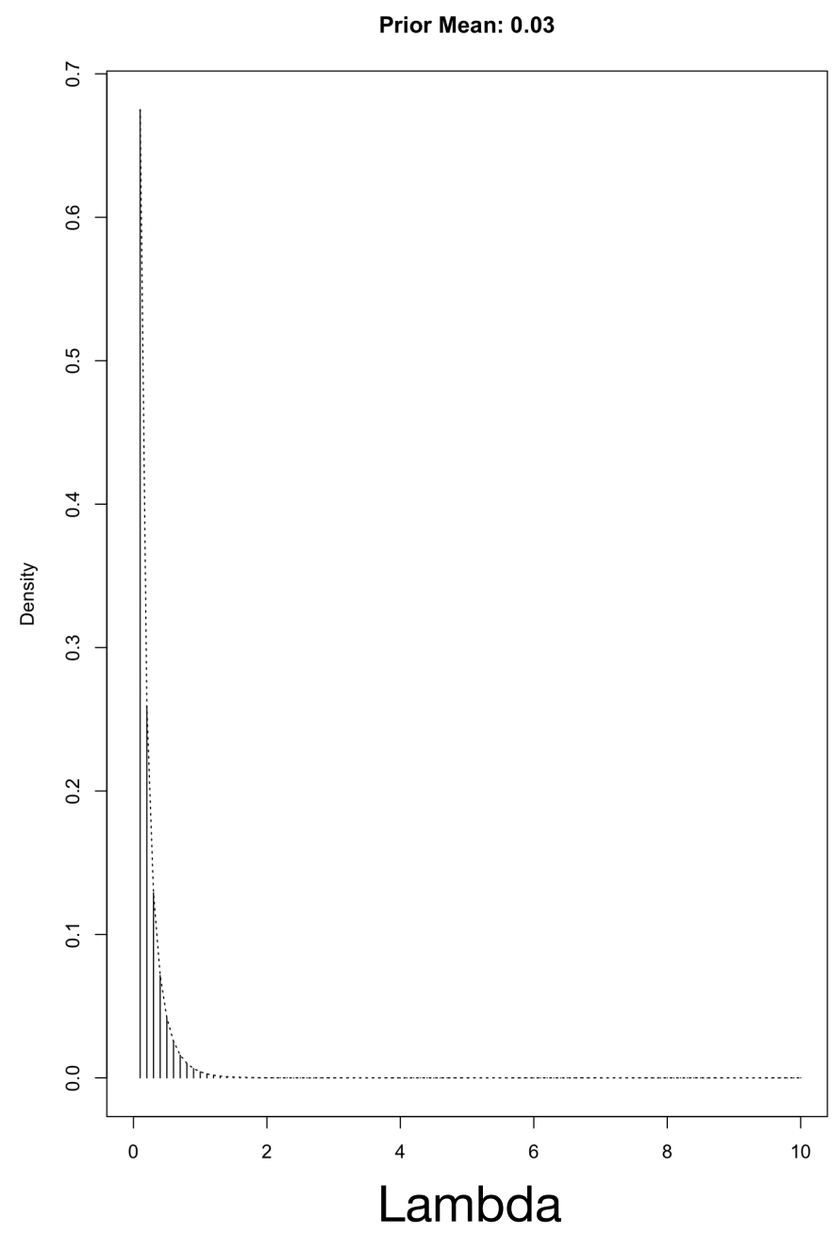
Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

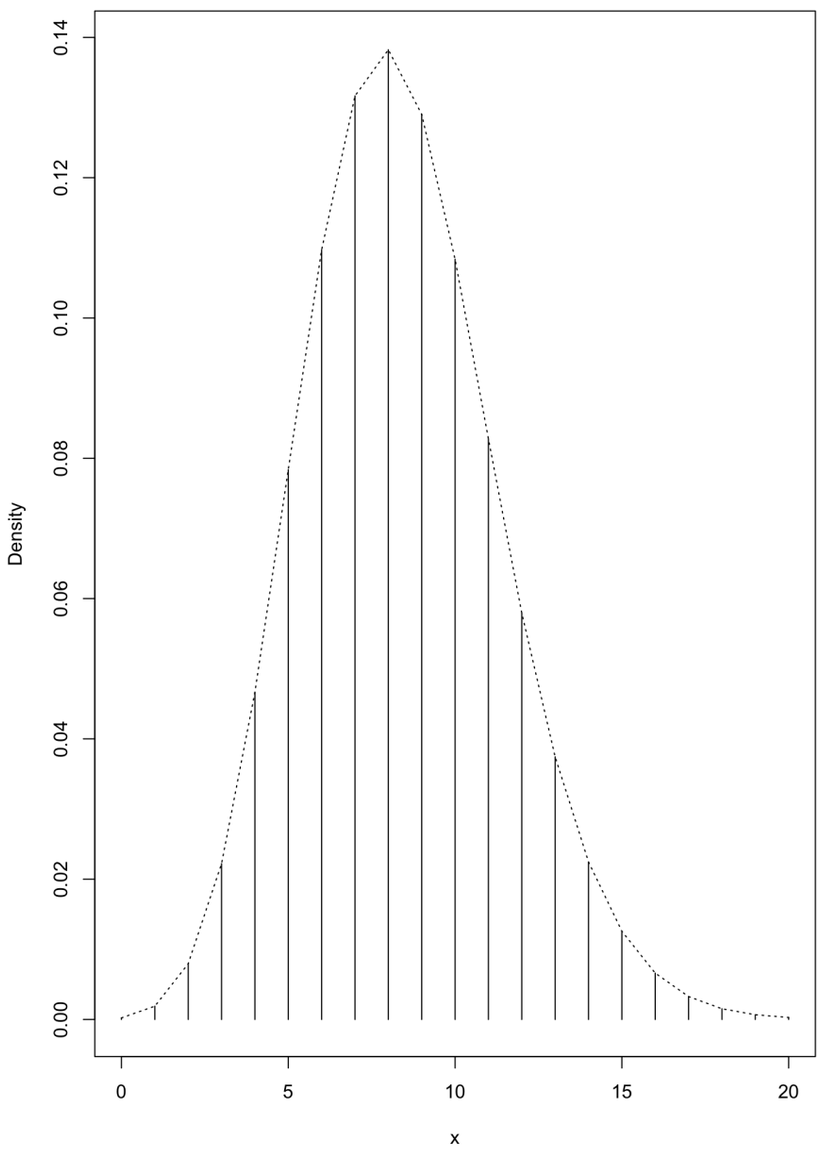


+

Information in the data

Likelihood

$$\hat{\lambda} = 8.4$$

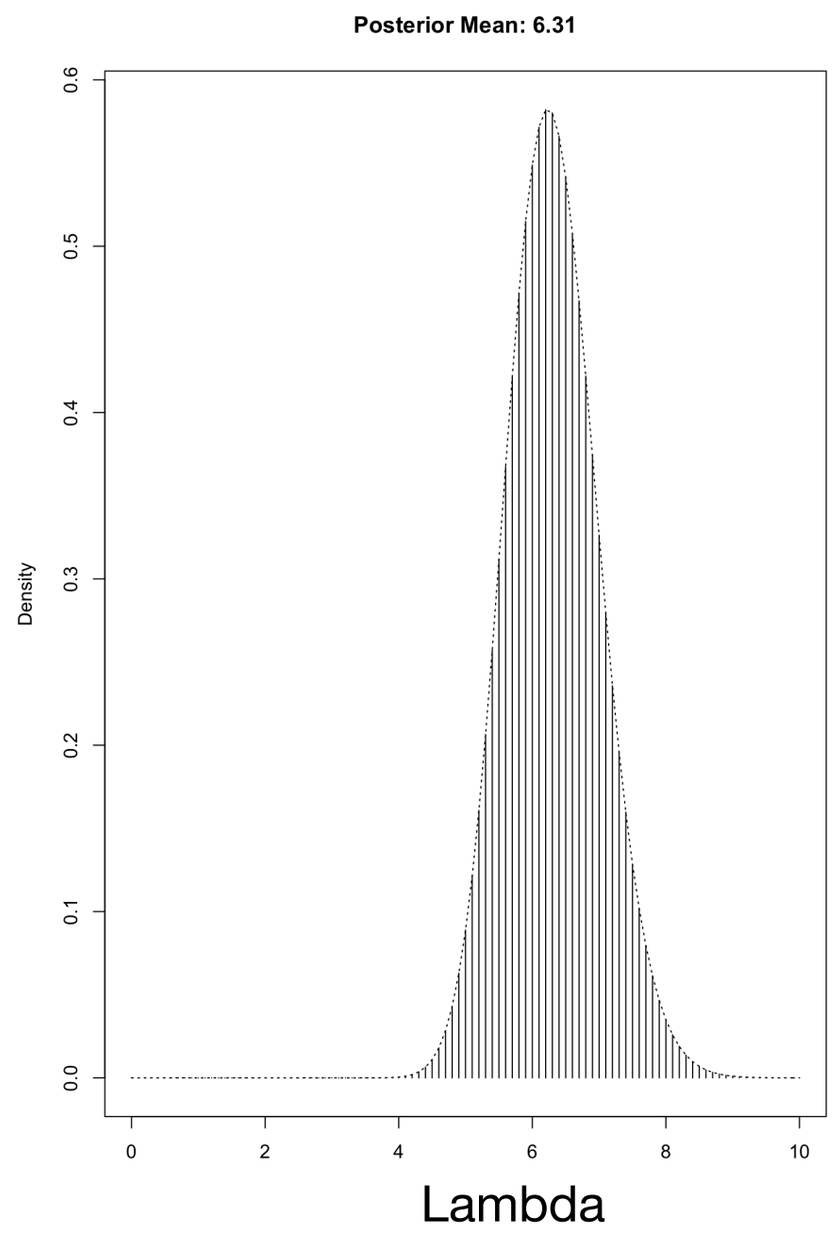


=

Inference on parameter

Posterior

$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

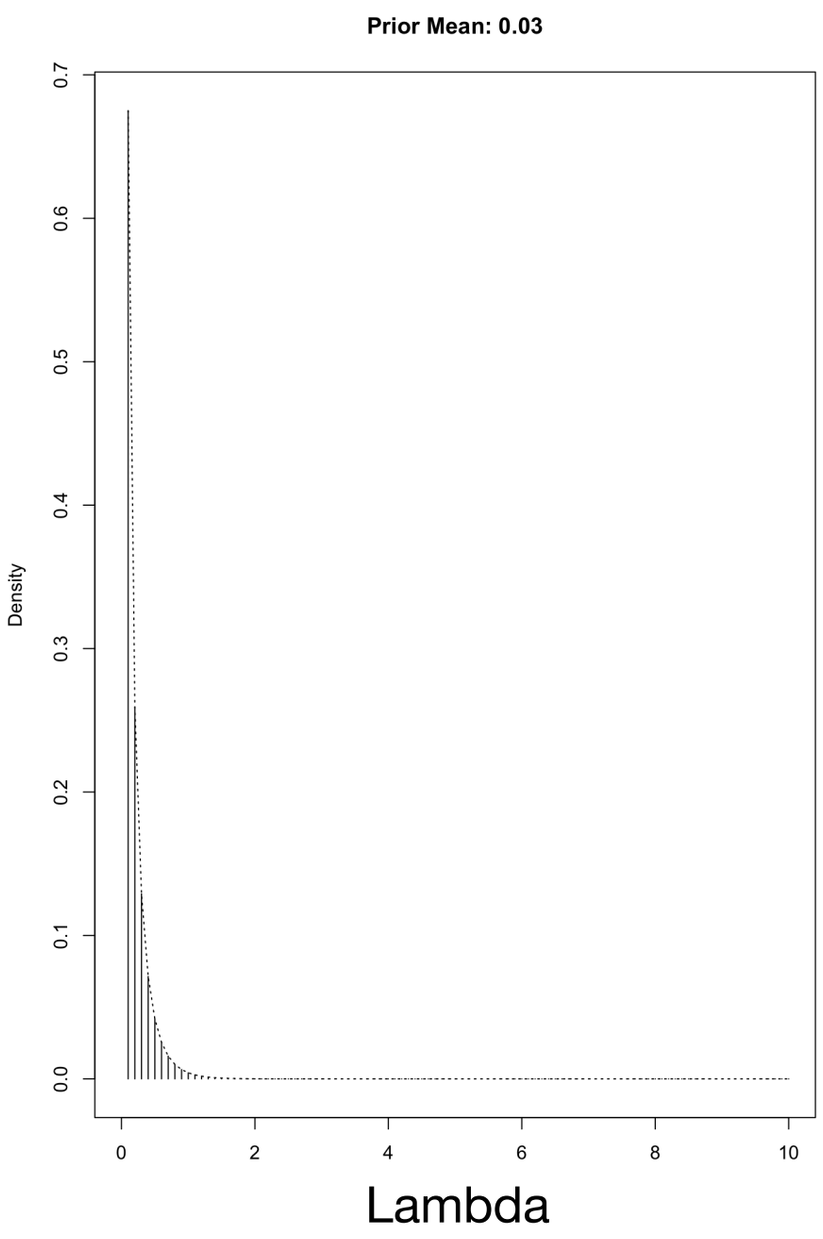


Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior



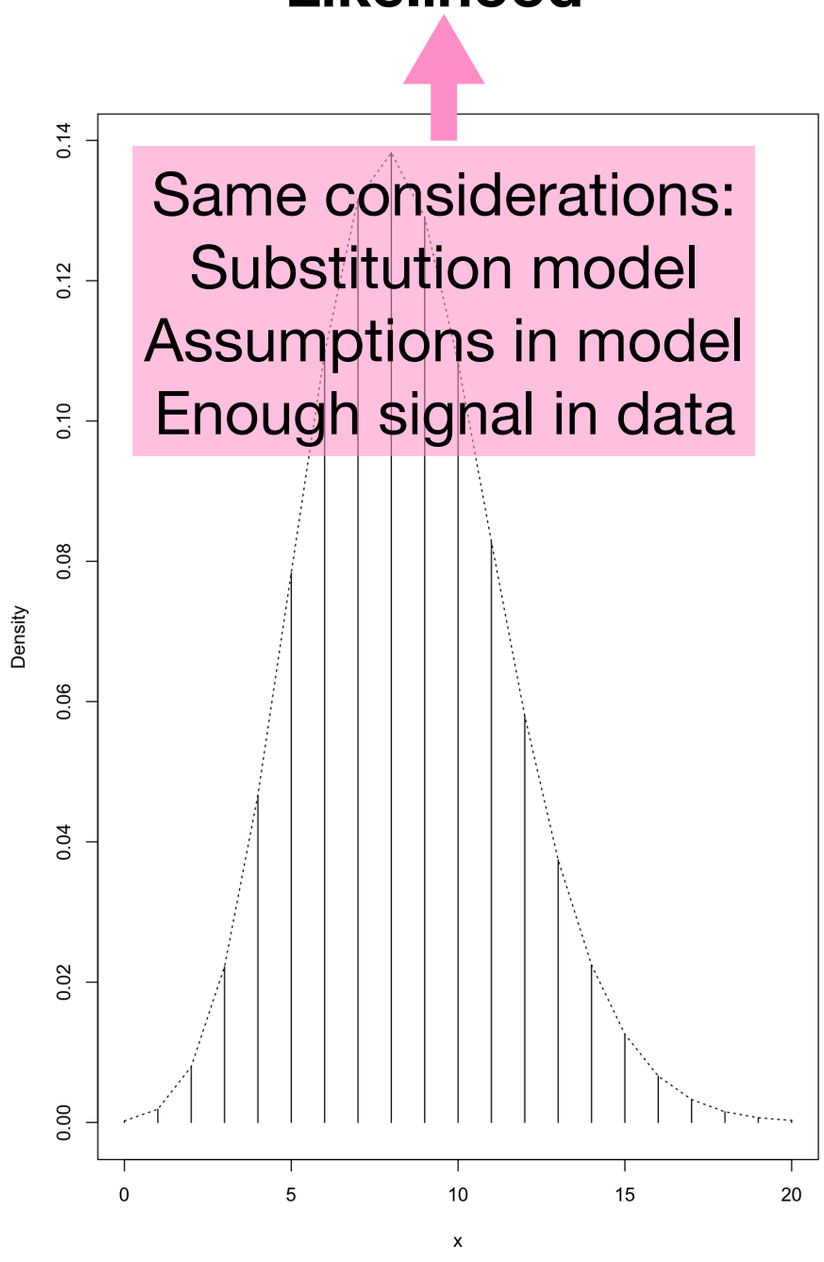
$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

+

Information in the data

Likelihood

Same considerations:
Substitution model
Assumptions in model
Enough signal in data

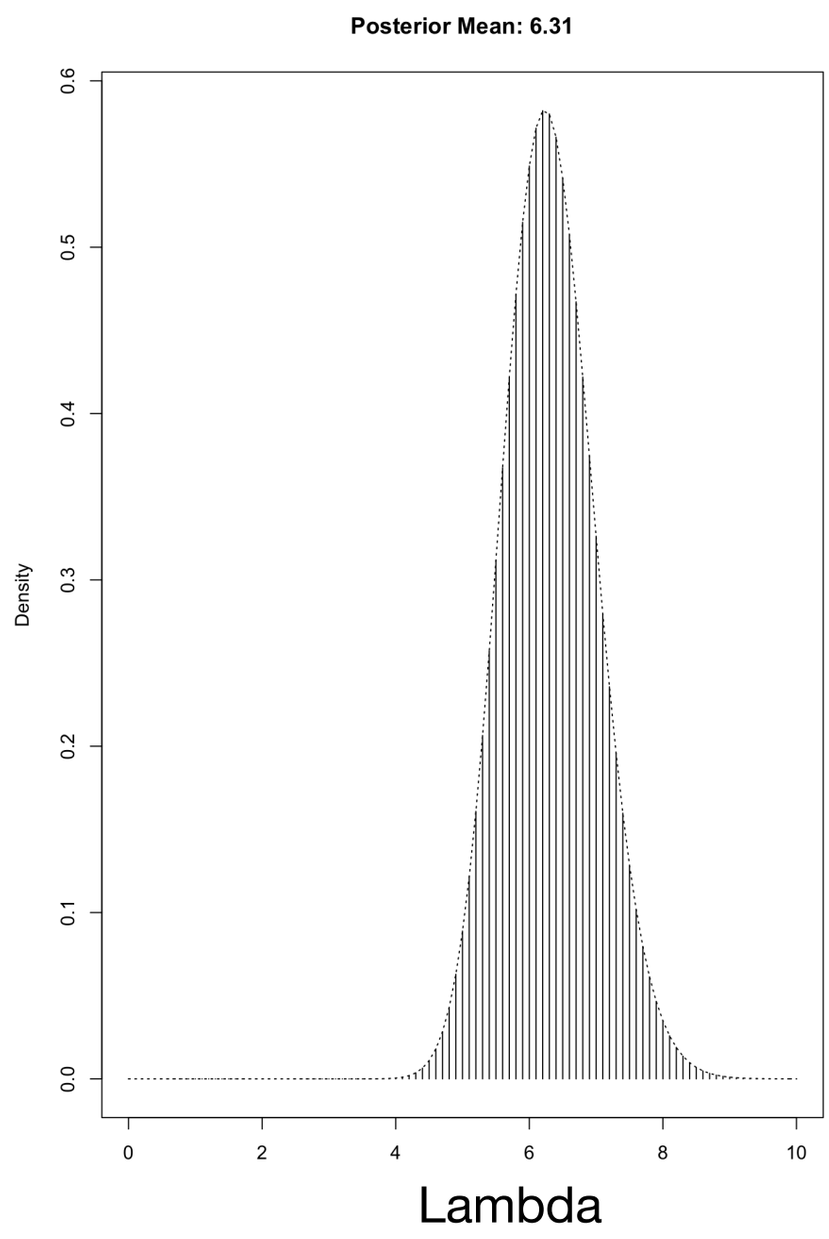


$$\hat{\lambda} = 8.4$$

=

Inference on parameter

Posterior



$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

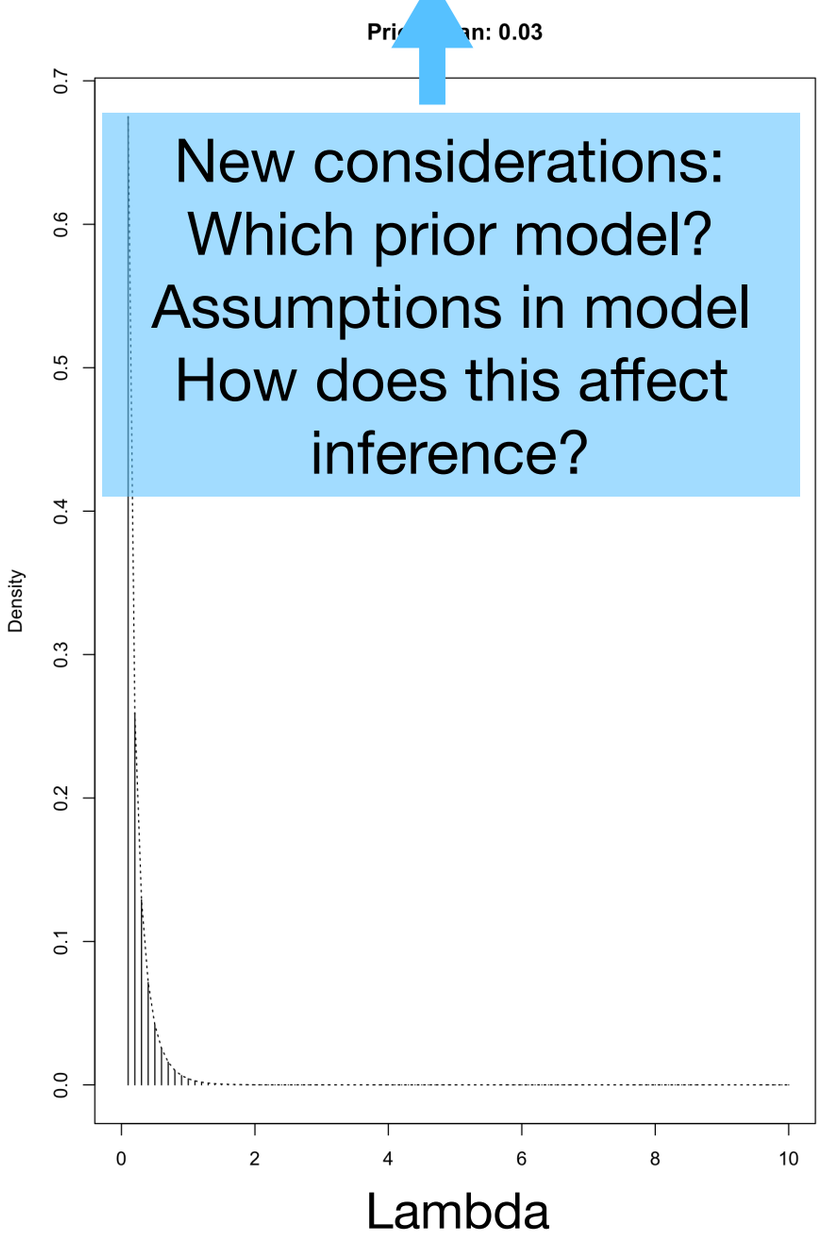
Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

New considerations:
Which prior model?
Assumptions in model
How does this affect inference?

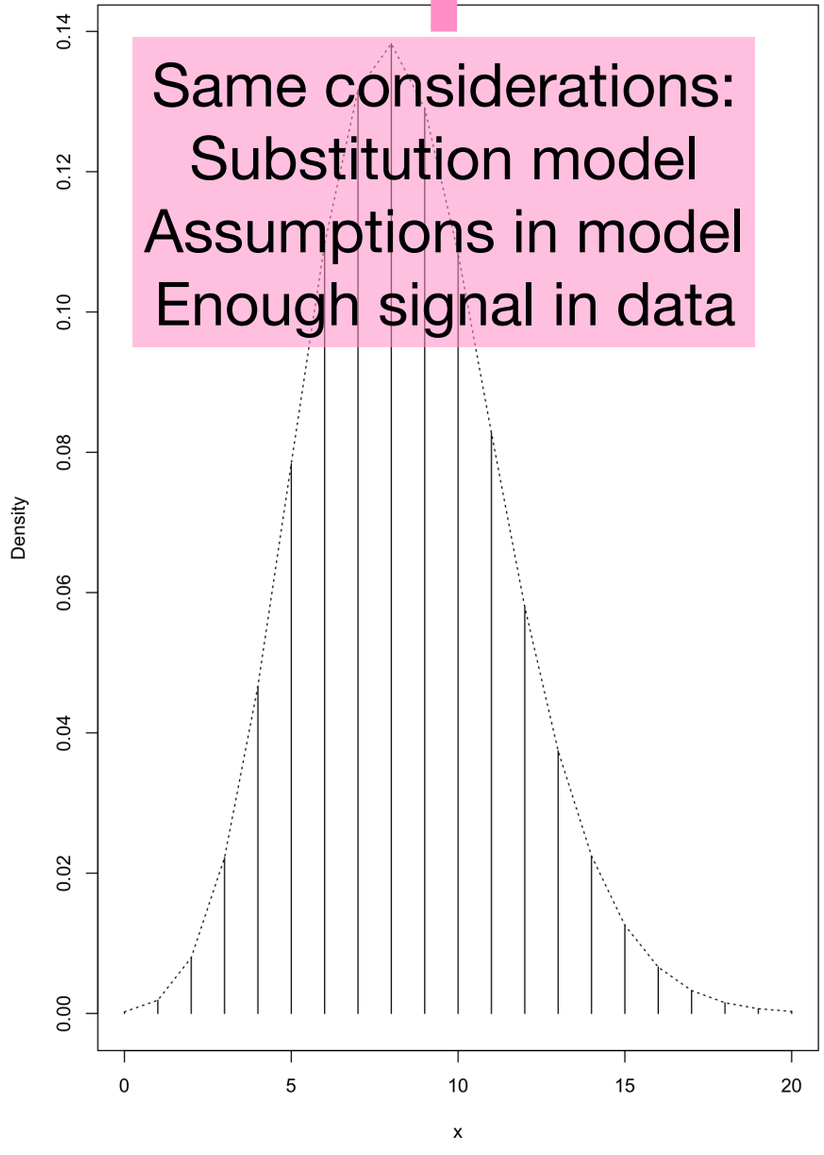


$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

Information in the data

Likelihood

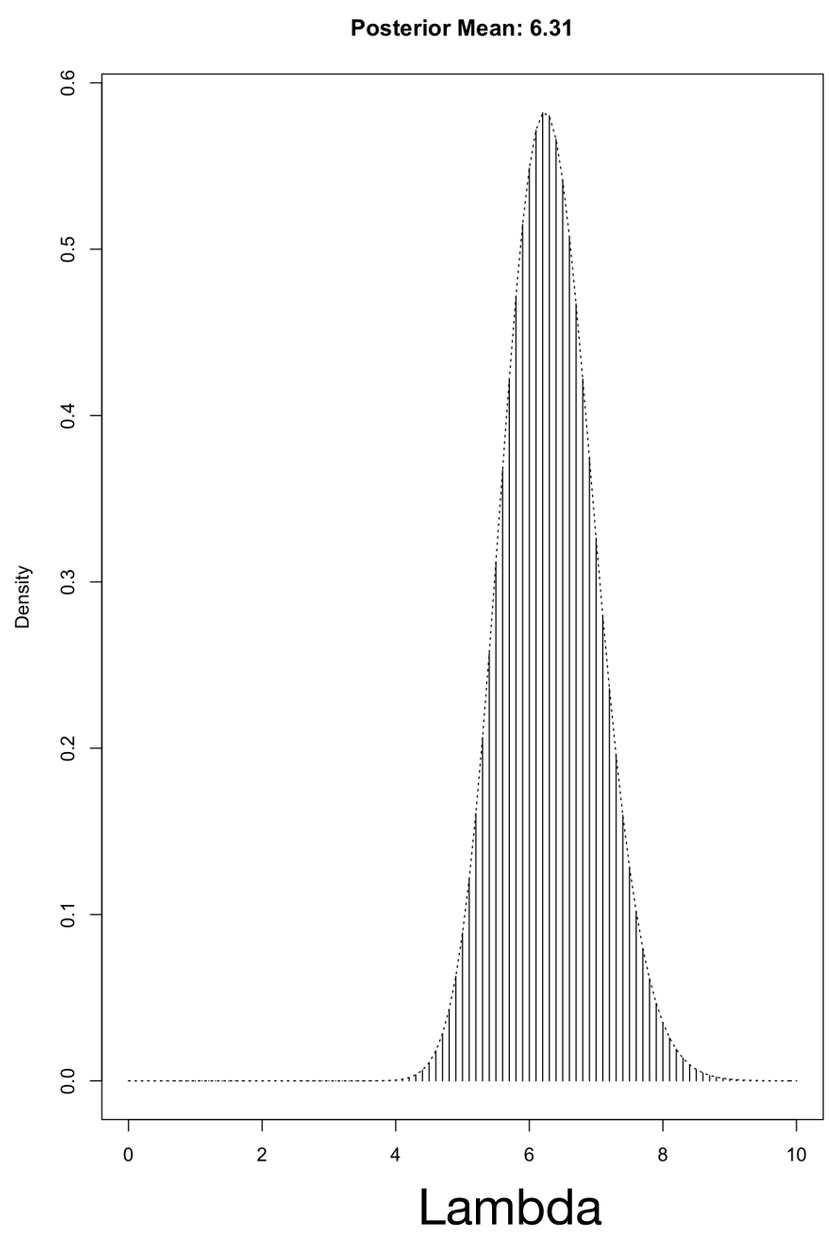
Same considerations:
Substitution model
Assumptions in model
Enough signal in data



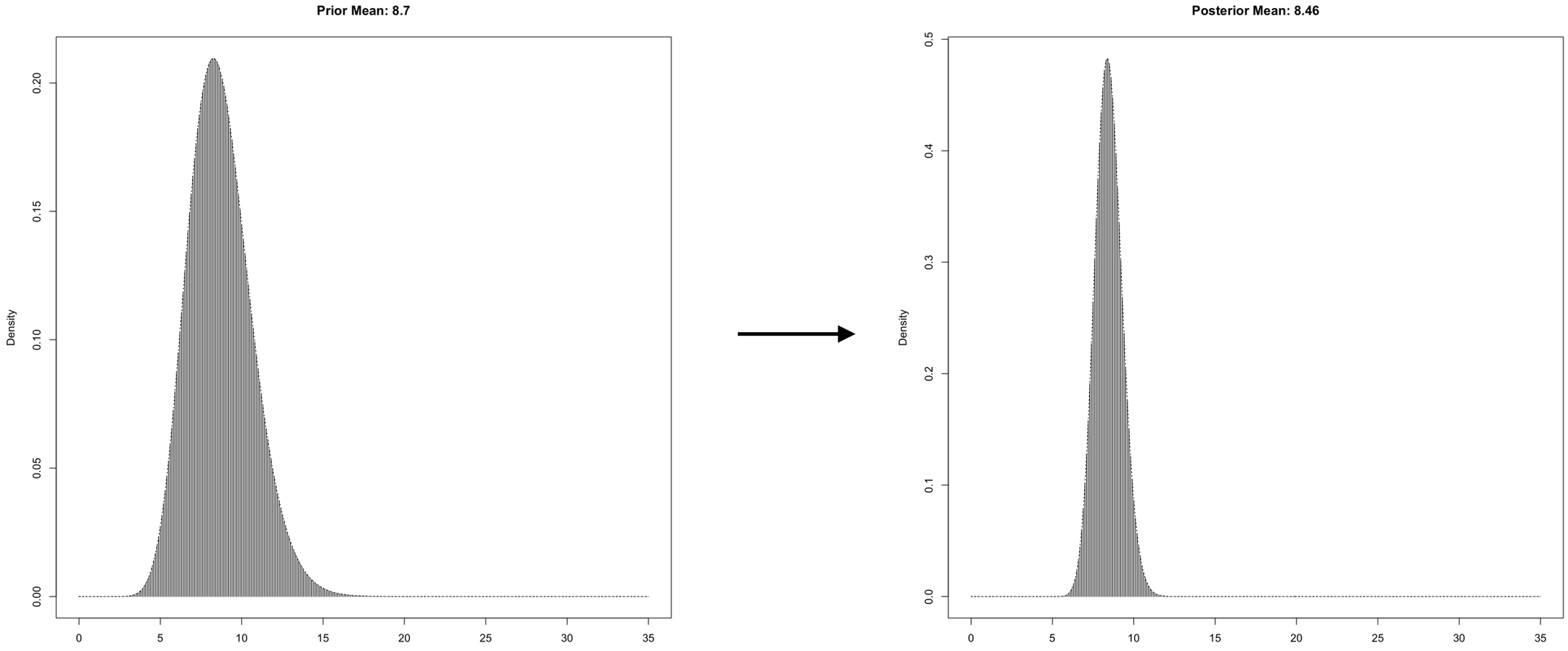
$$\hat{\lambda} = 8.4$$

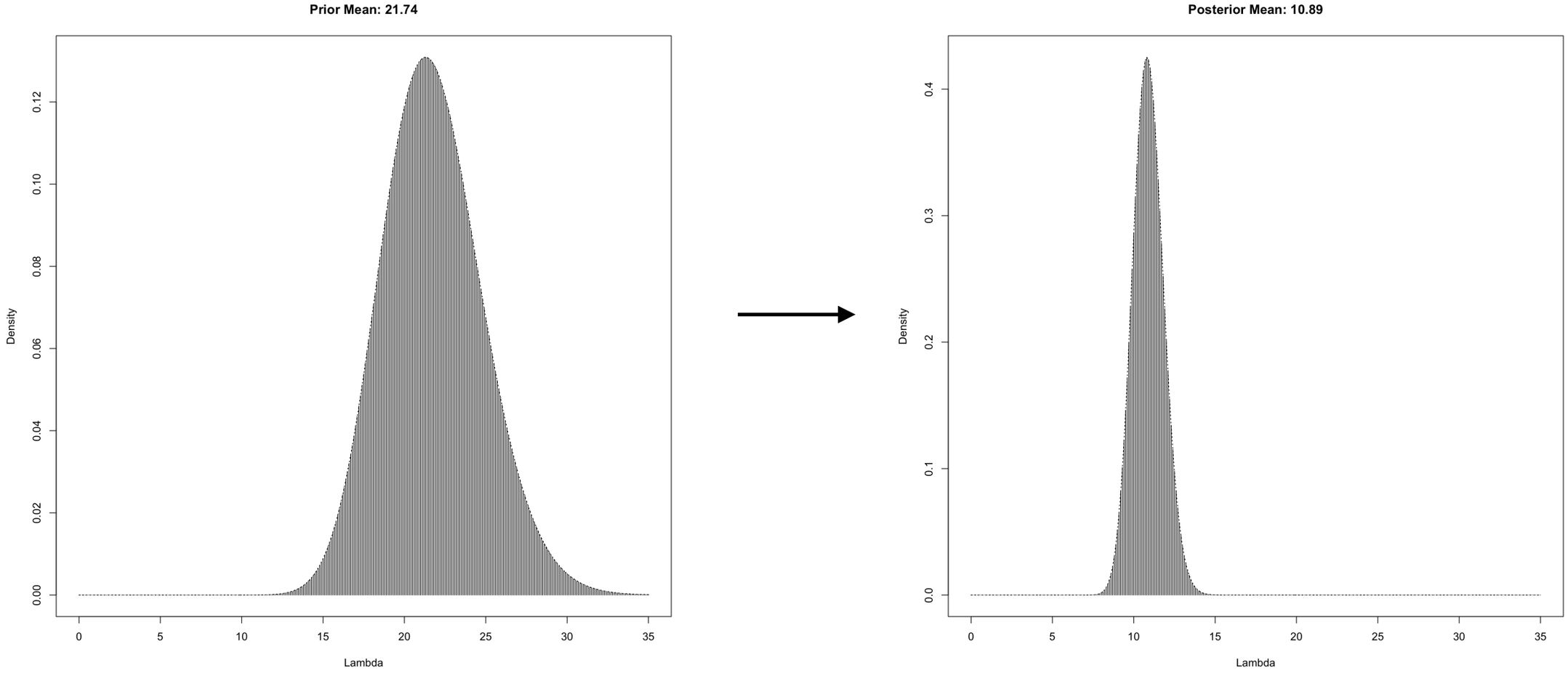
Inference on parameter

Posterior

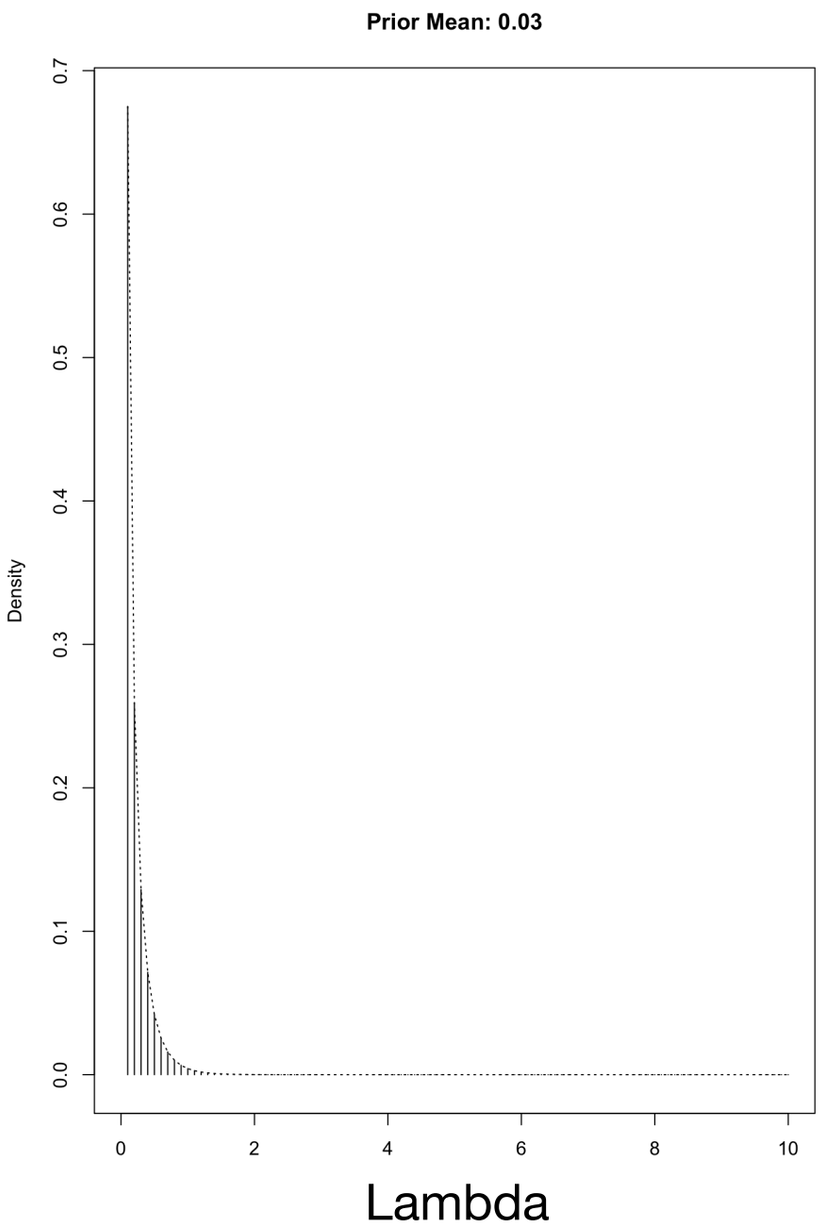


$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$


Prior Mean: 8.7
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
35
Posterior Mean: 8.46
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
30
35

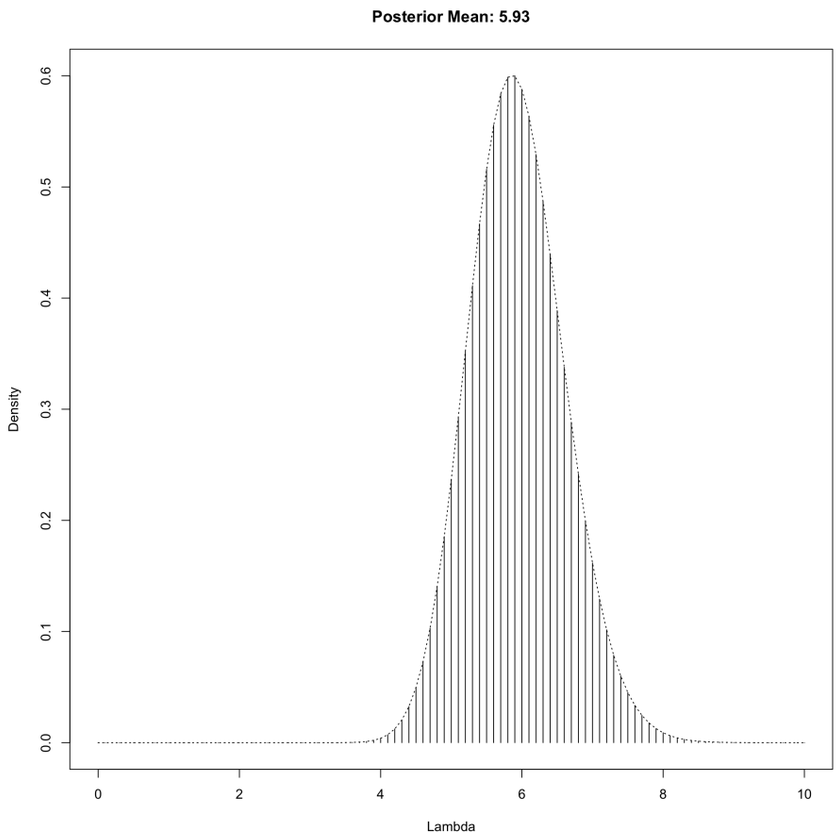

Prior Mean: 21.74
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
25
30
35
Lambda
Posterior Mean: 10.89
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
30
35
Lambda

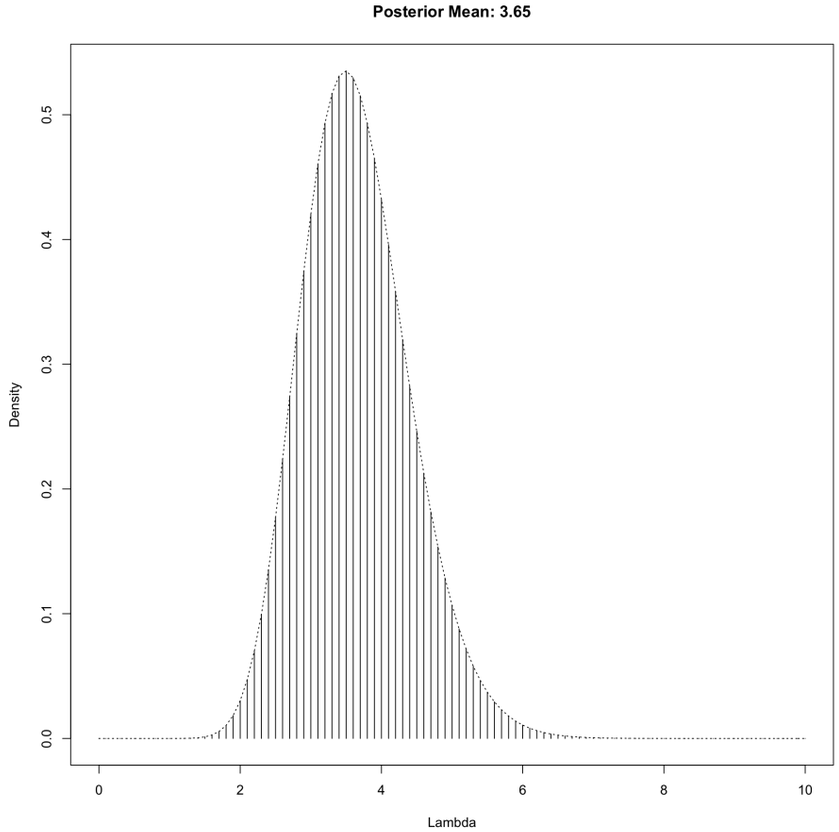
Prior



$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

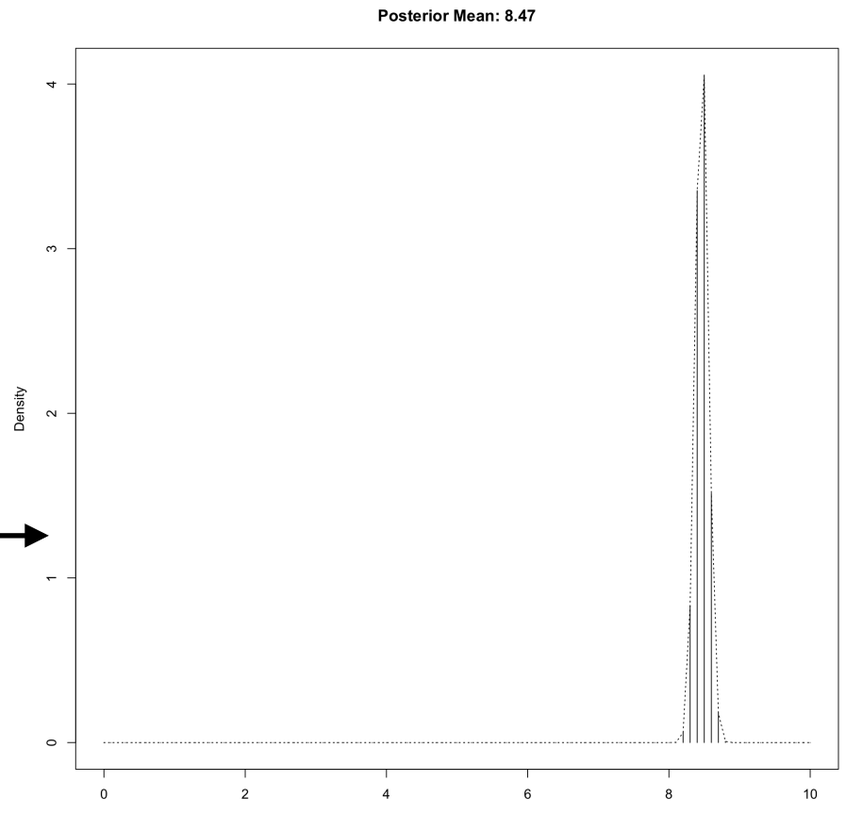
n=3 →





← **n=10**

n=1000 →



Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

New considerations:
Which prior model?
Assumptions in model
How does this affect inference?

- Pulls the posterior towards it
- With sufficient enough data, it does not matter
- With non-informative/ flat priors, you get the same as MLE

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

+

Information in the data

Likelihood

Same considerations:
Substitution model
Assumptions in model
Enough signal in data

$$\hat{\lambda} = 8.4$$

=

Inference on parameter

Posterior

$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\textbf{Likelihood}\ \prod P(X = x_i|\lambda)\ \textbf{Prior}\ p(\lambda)}{p(\mathbf{X})\ \textbf{Marginal}}$$

← Many times, intractable

Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\textbf{Likelihood}\ \prod P(X = x_i|\lambda)\ \textbf{Prior}\ p(\lambda)}{p(\mathbf{X})\ \textbf{Marginal}}$$

← Many times, intractable

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \left(\frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}\right) \lambda^{\alpha-1}e^{-\beta\lambda}$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \left(\frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}\right) \lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \left(\frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}\right) \lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda} \leftarrow Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \left(\frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}\right) \lambda^{\alpha-1}e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda} \leftarrow Gamma\left(\sum x_i + \alpha, n+\beta\right)$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i + \alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i+\alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i+\alpha, n+\beta\right)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$ **Conjugate prior**

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i+\alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n+\beta\right)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$ **Conjugate prior**

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i+\alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n+\beta\right)$$

Posterior mean: $\frac{\sum x_i+\alpha}{n+\beta} = \frac{n}{n+\beta}\left(\frac{\sum x_i}{n}\right) + \frac{\beta}{n+\beta}\left(\frac{\alpha}{\beta}\right)$

Intractable posterior distribution

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2} \right)$$

Intractable posterior distribution

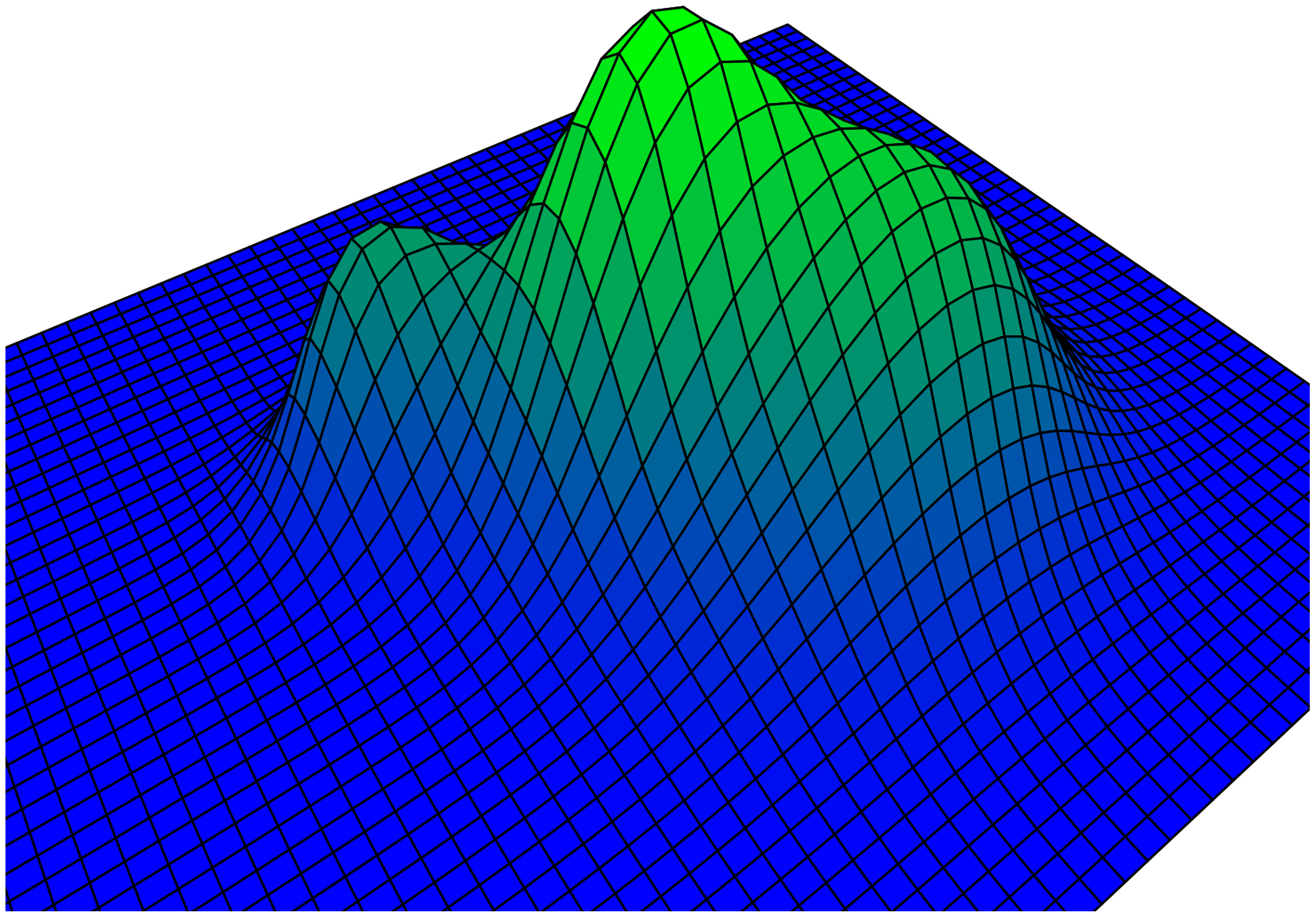
$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Not a known distribution anymore We need to approximate it

MCMC: A way to approximate intractable posterior distributions



MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution
Usually symmetric, but could be asymmetric (Hasting ratio)

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution
Usually symmetric, but could be asymmetric (Hasting ratio)

Acceptance probability
Posterior ratio

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$

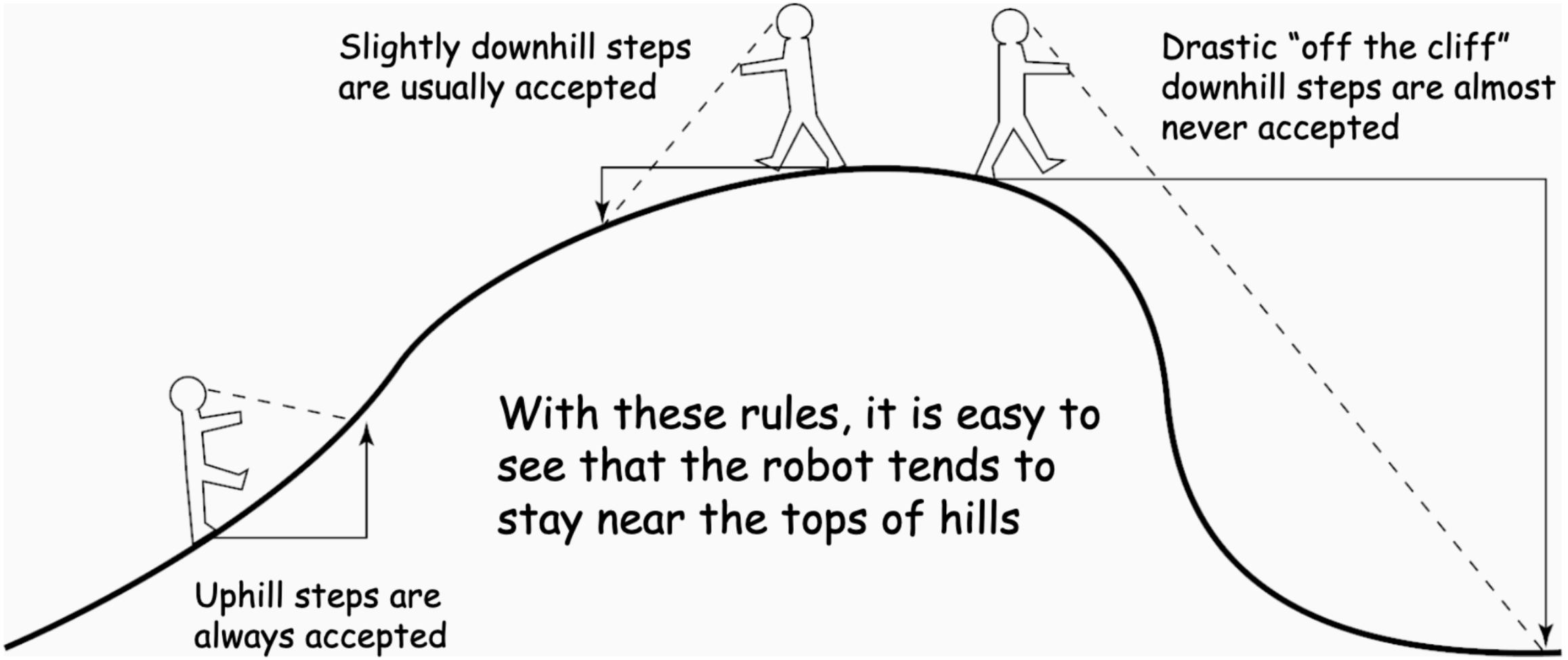
- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution
Usually symmetric, but could be asymmetric (Hasting ratio)

Acceptance probability
Posterior ratio

The chain tends to stay on regions of high posterior

MCMC robot's rules



Hasting ratio

$$\alpha = \left[\frac{P(\lambda^*|\mathbf{X})}{P(\lambda|\mathbf{X})}\right]\left[\frac{q(\lambda|\lambda^*)}{q(\lambda^*|\lambda)}\right]$$

Proposal distribution

Posterior ratio **Hasting ratio**

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Optional (highly recommended) homework:

- Code the MCMC for this example in your preferred programming language
- Plot the MCMC histogram and compute the posterior mean for lambda
- Play with different values for w and n (length of the chain)

MCMC considerations

- Choice of priors
- Mixing
- Convergence
- Burnin

MCMC considerations

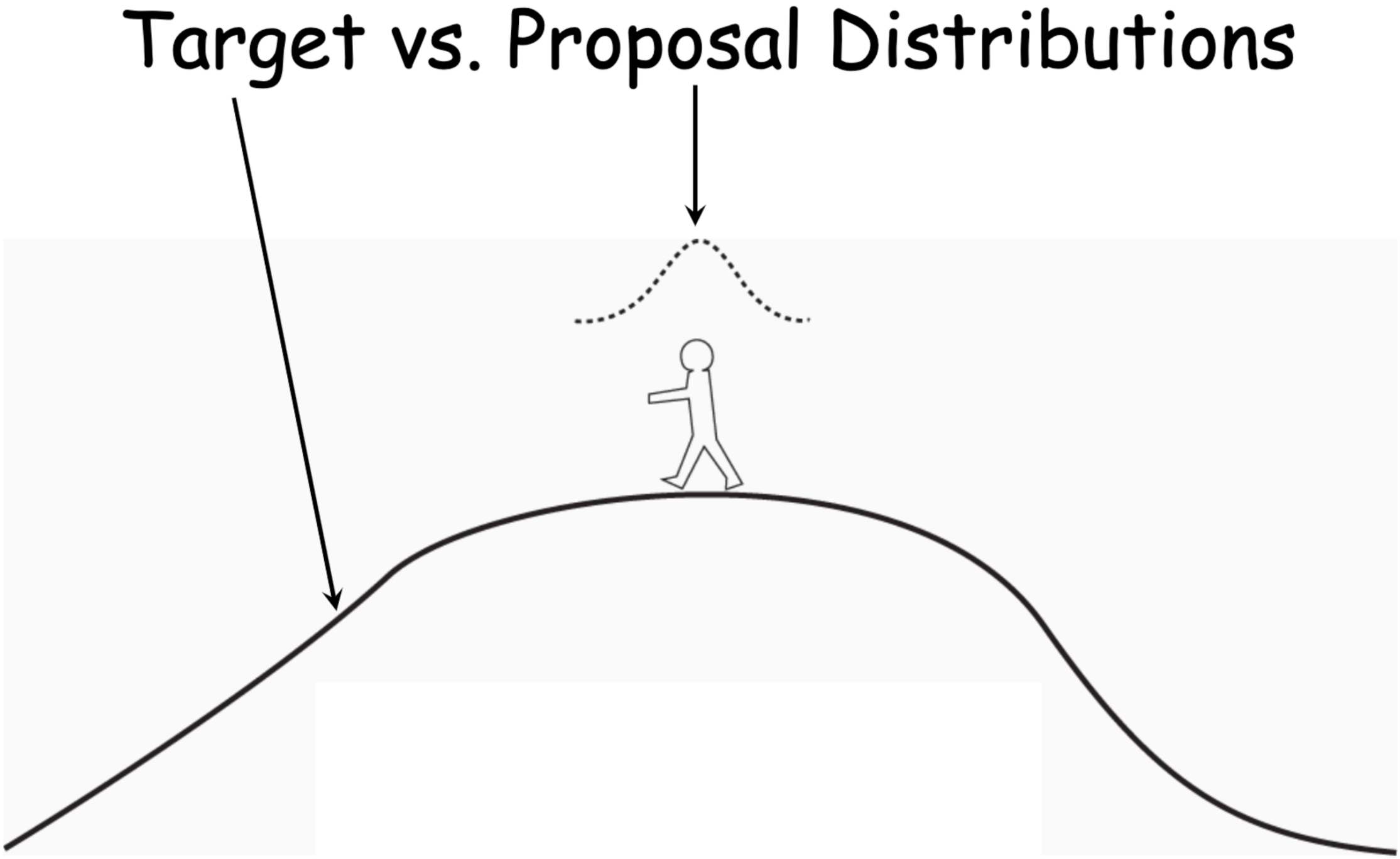
- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence
- Burnin

MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin

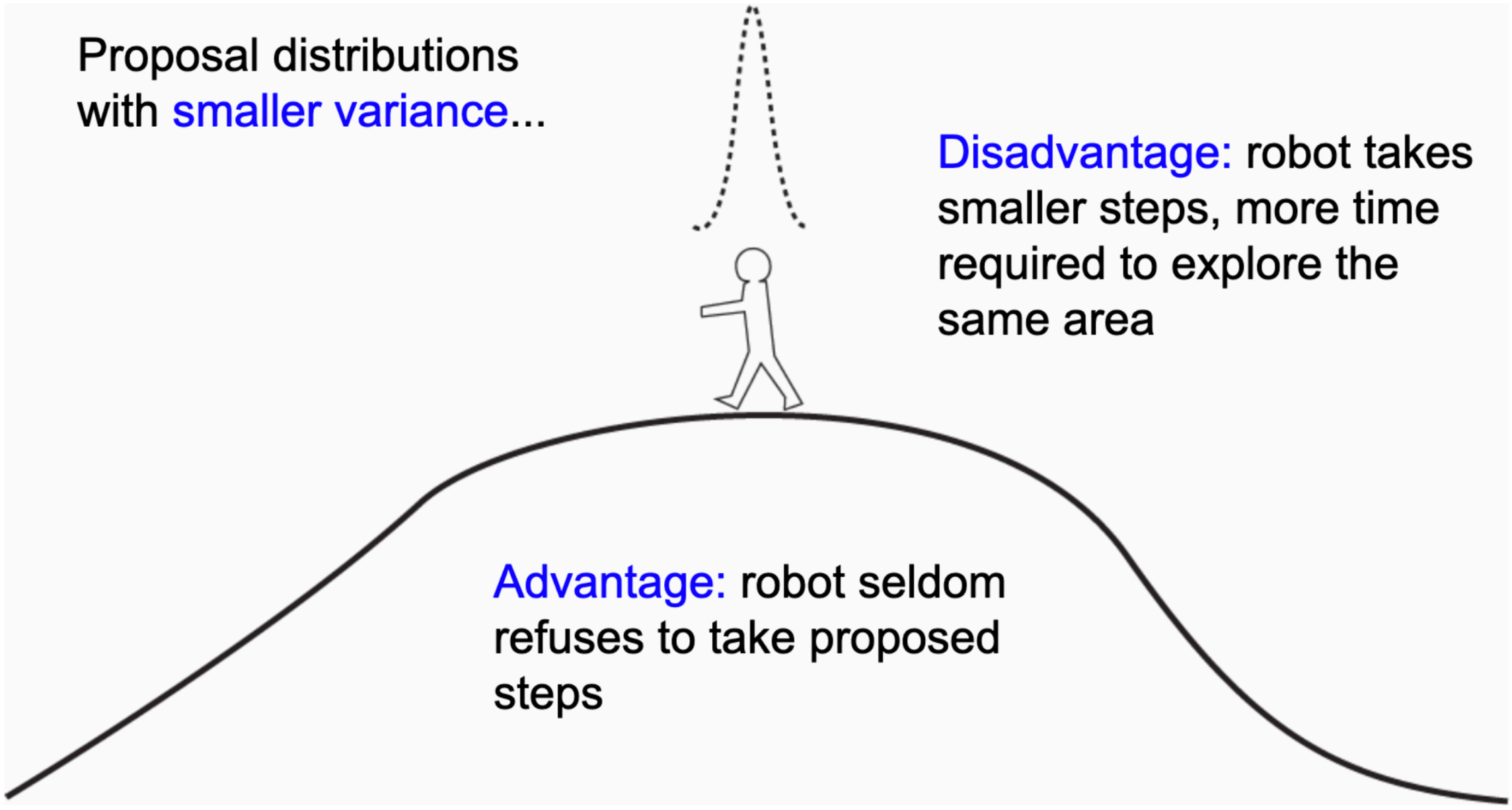
MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin ← How long it takes to reach regions of high posterior values?



Target vs. Proposal Distributions

Proposal distributions with smaller variance...

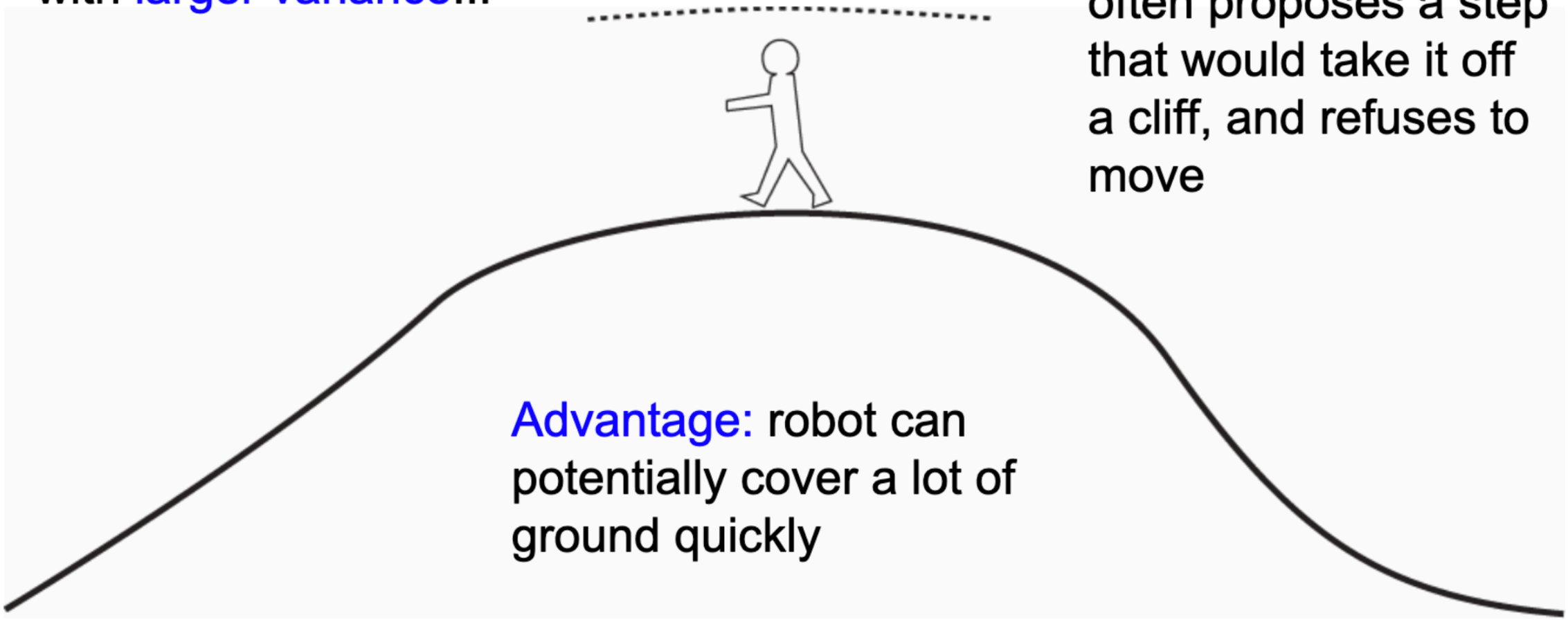


Disadvantage: robot takes smaller steps, more time required to explore the same area

Advantage: robot seldom refuses to take proposed steps

Target vs. Proposal Distributions

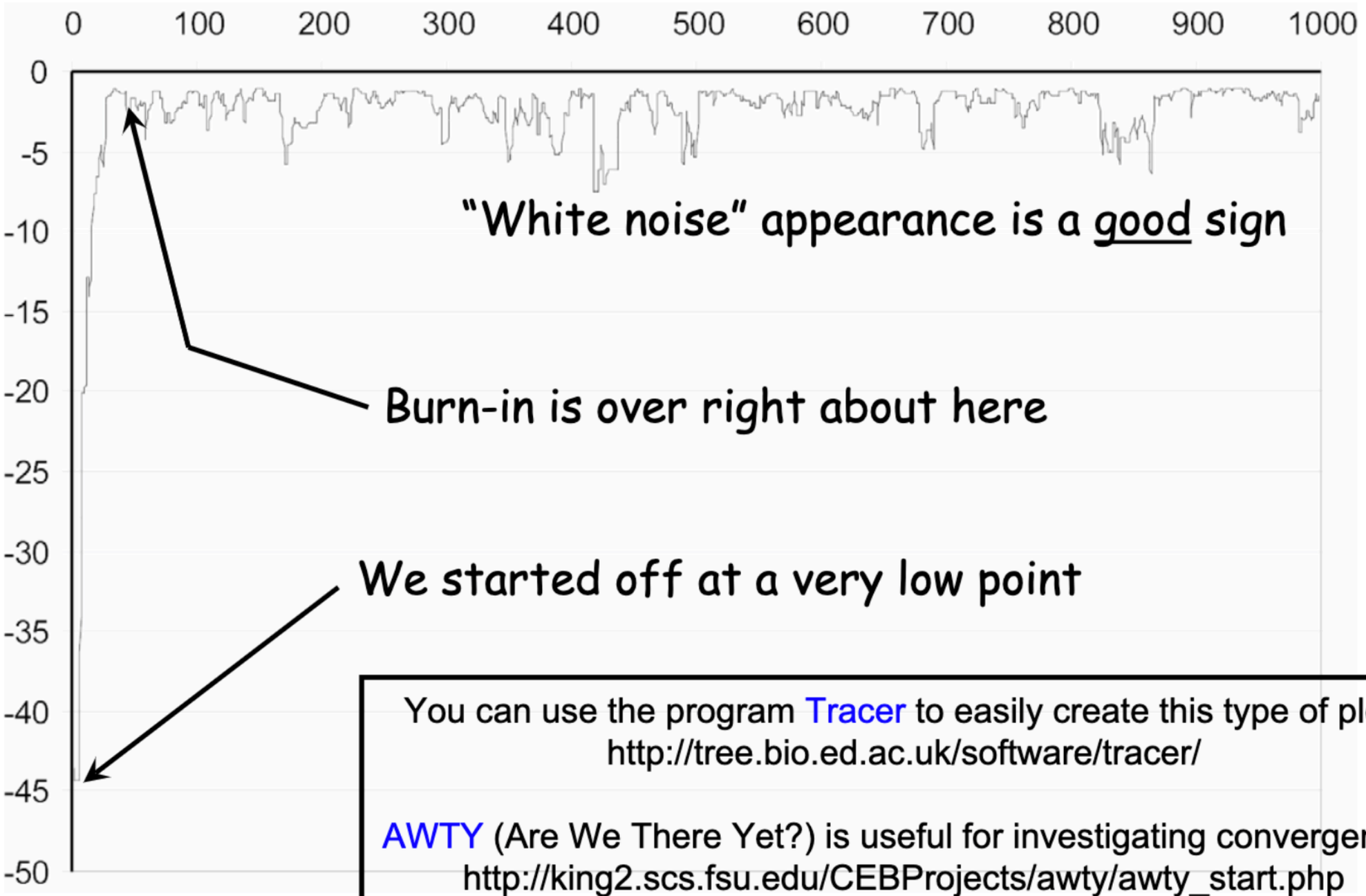
Proposal distributions with larger variance...



Disadvantage: robot often proposes a step that would take it off a cliff, and refuses to move

Advantage: robot can potentially cover a lot of ground quickly

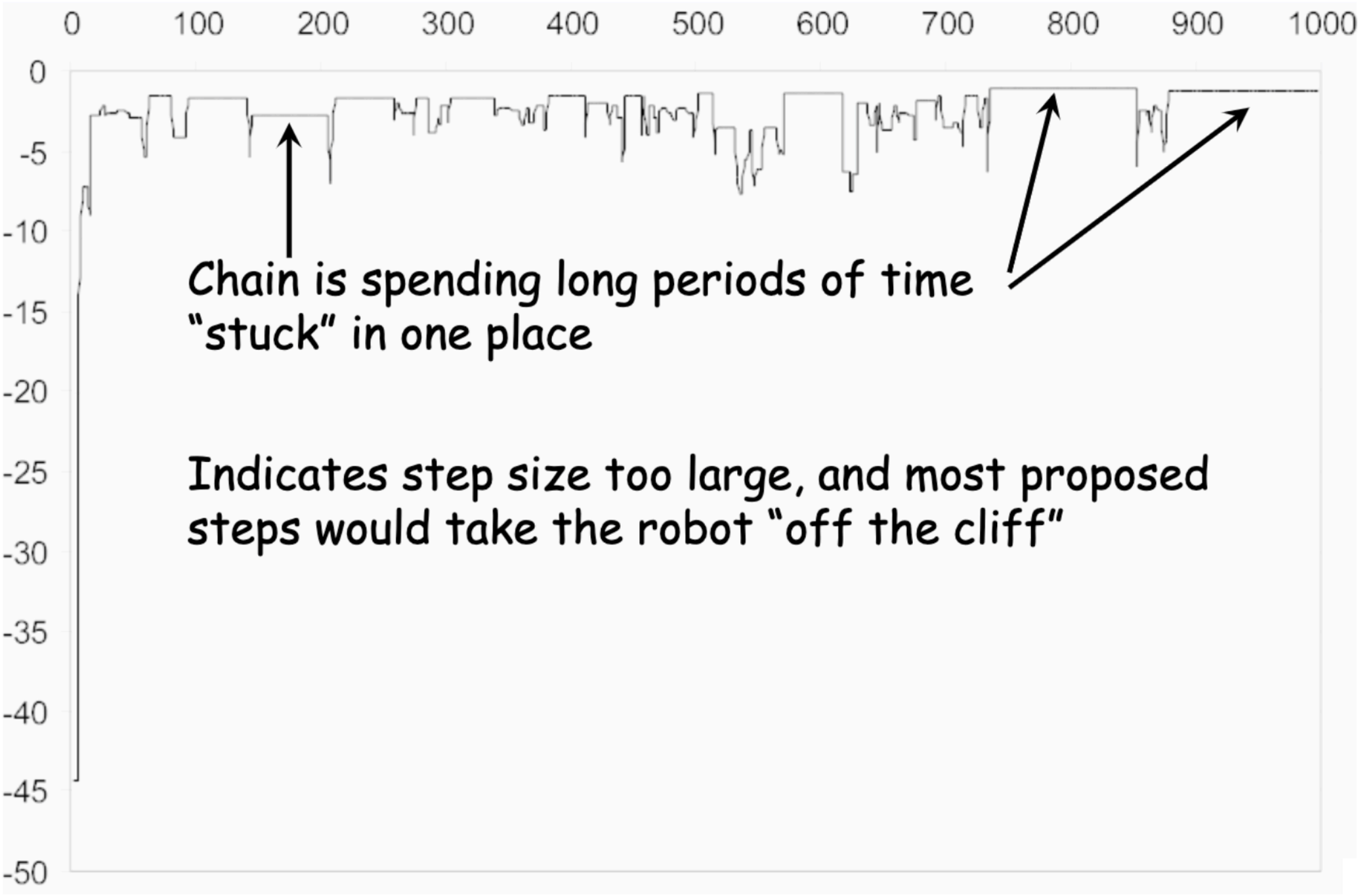
Trace plots



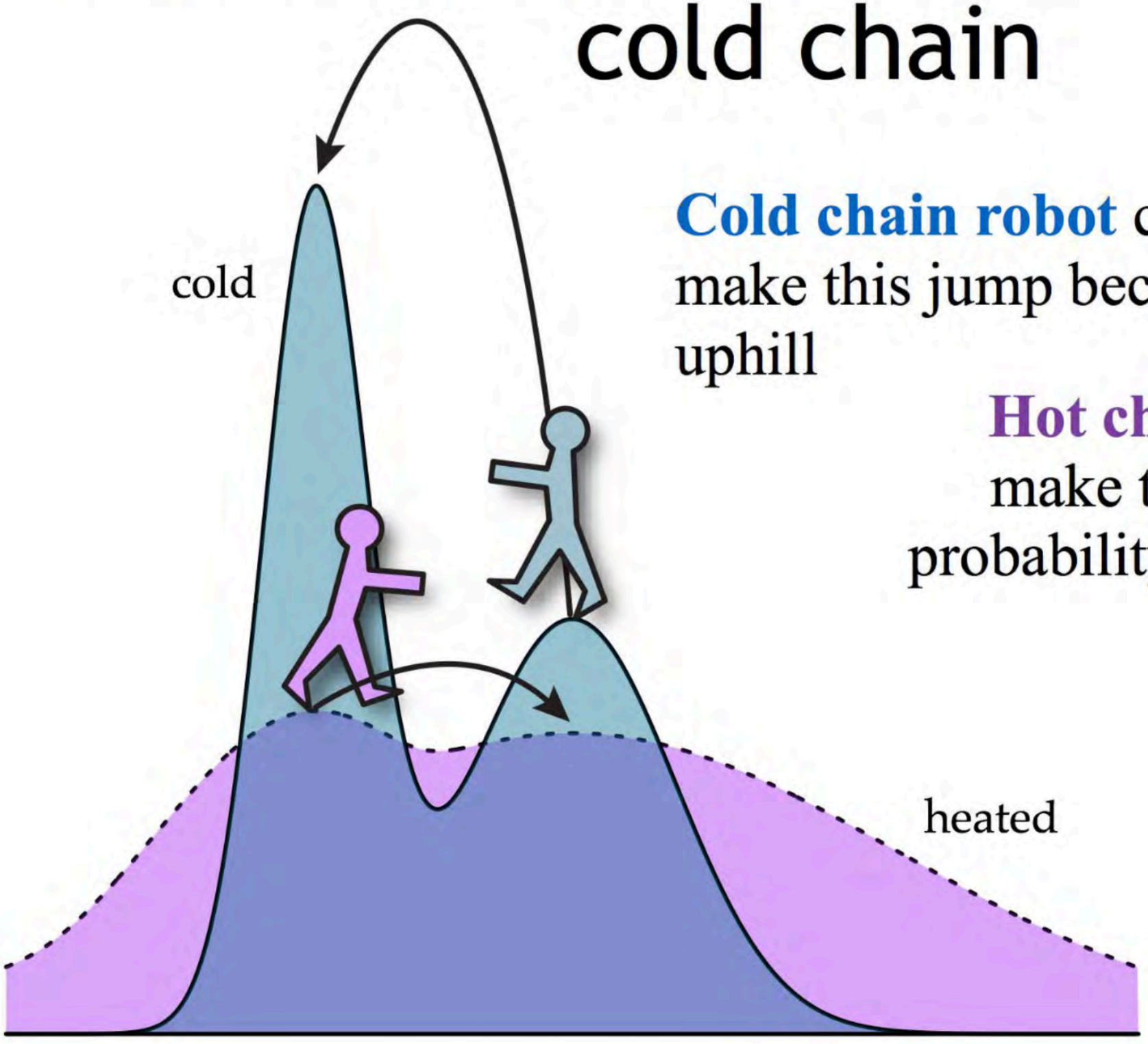
You can use the program Tracer to easily create this type of plot:
http://tree.bio.ed.ac.uk/software/tracer/

AWTY (Are We There Yet?) is useful for investigating convergence:
http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php

Poor mixing



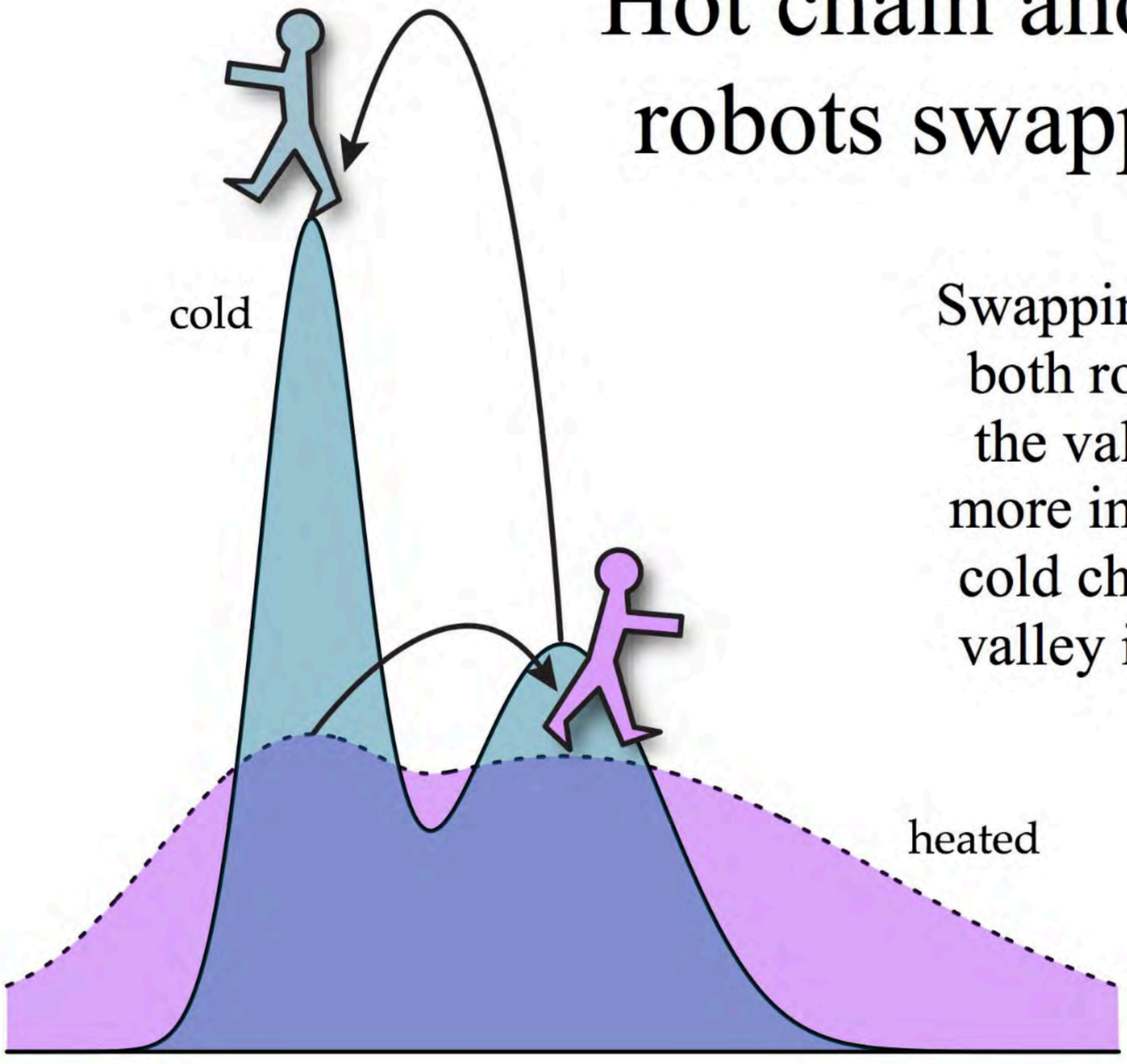
Heated chains act as scouts for the cold chain



Cold chain robot can easily make this jump because it is uphill

Hot chain robot can also make this jump with high probability because it is only slightly downhill

Hot chain and cold chain robots swapping places



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

"Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough". - Paul Lewis

"People always forget how long of a time infinity really is" - paraphrased from Dave Swofford

HW reading and quiz

- Nascimento et al (2017)

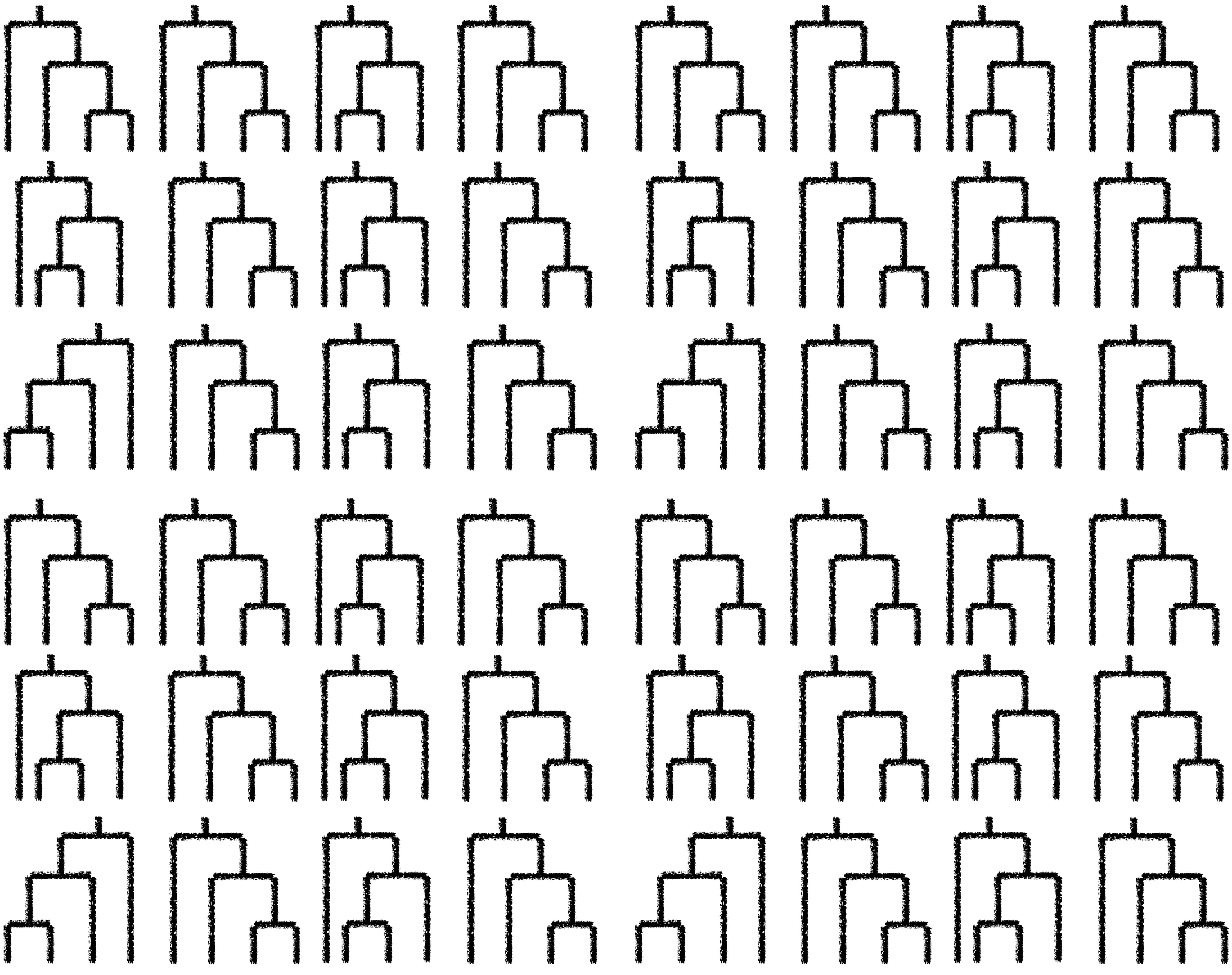
- Deadline: April 1, 2022 2:30pm CT

Back to phylogenetics

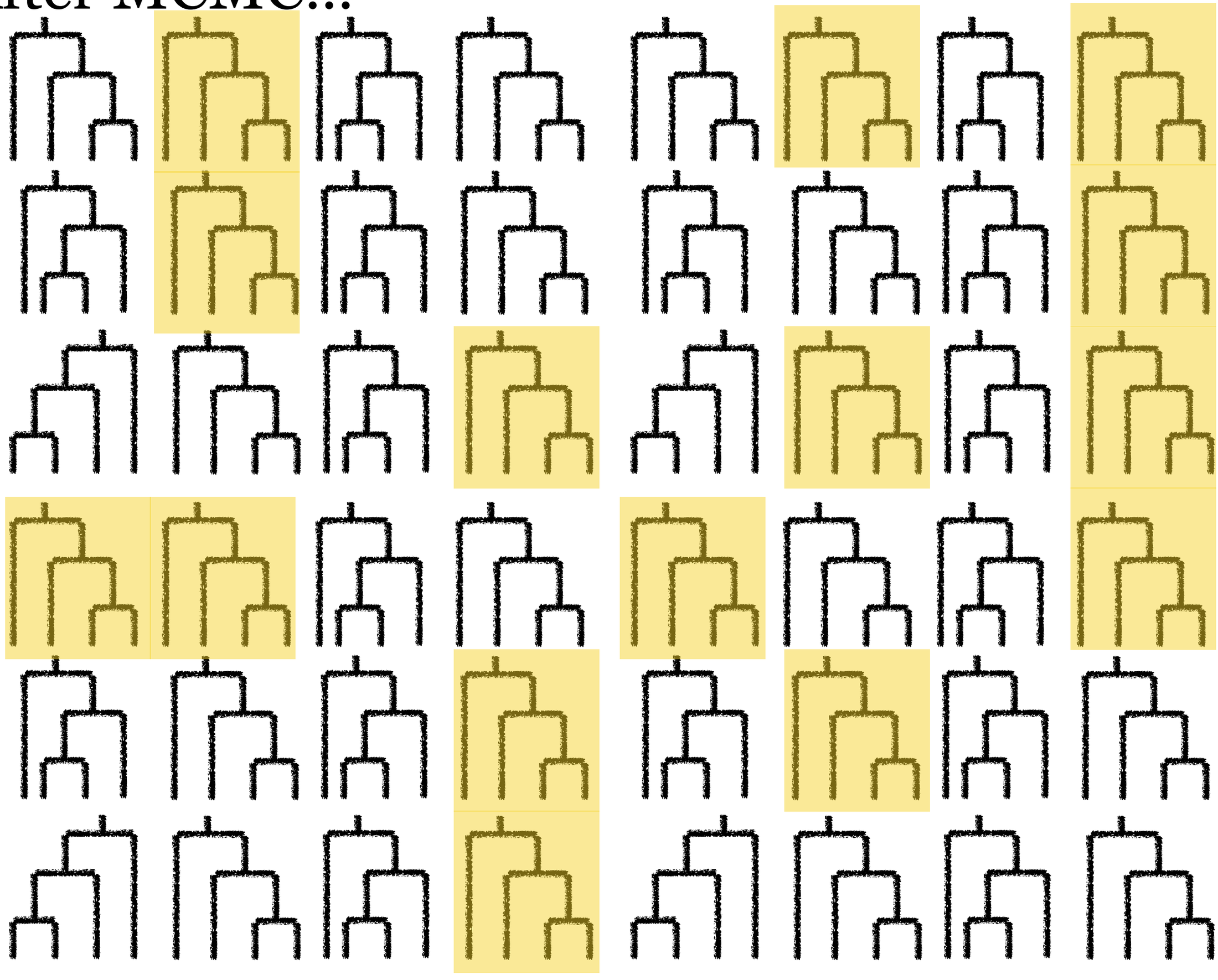
- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
  - Propose a **new tree** (e.g. Larget-Simon move) and either accept or reject the move
  - Propose (and either accept or reject) a **new model parameter value**
- Every *k* generations, save tree topology, branch lengths and all model parameters (i.e. **sample the chain**)
- After *n* generations, **summarize sample** using histograms, means, credible intervals, etc.

After MCMC…

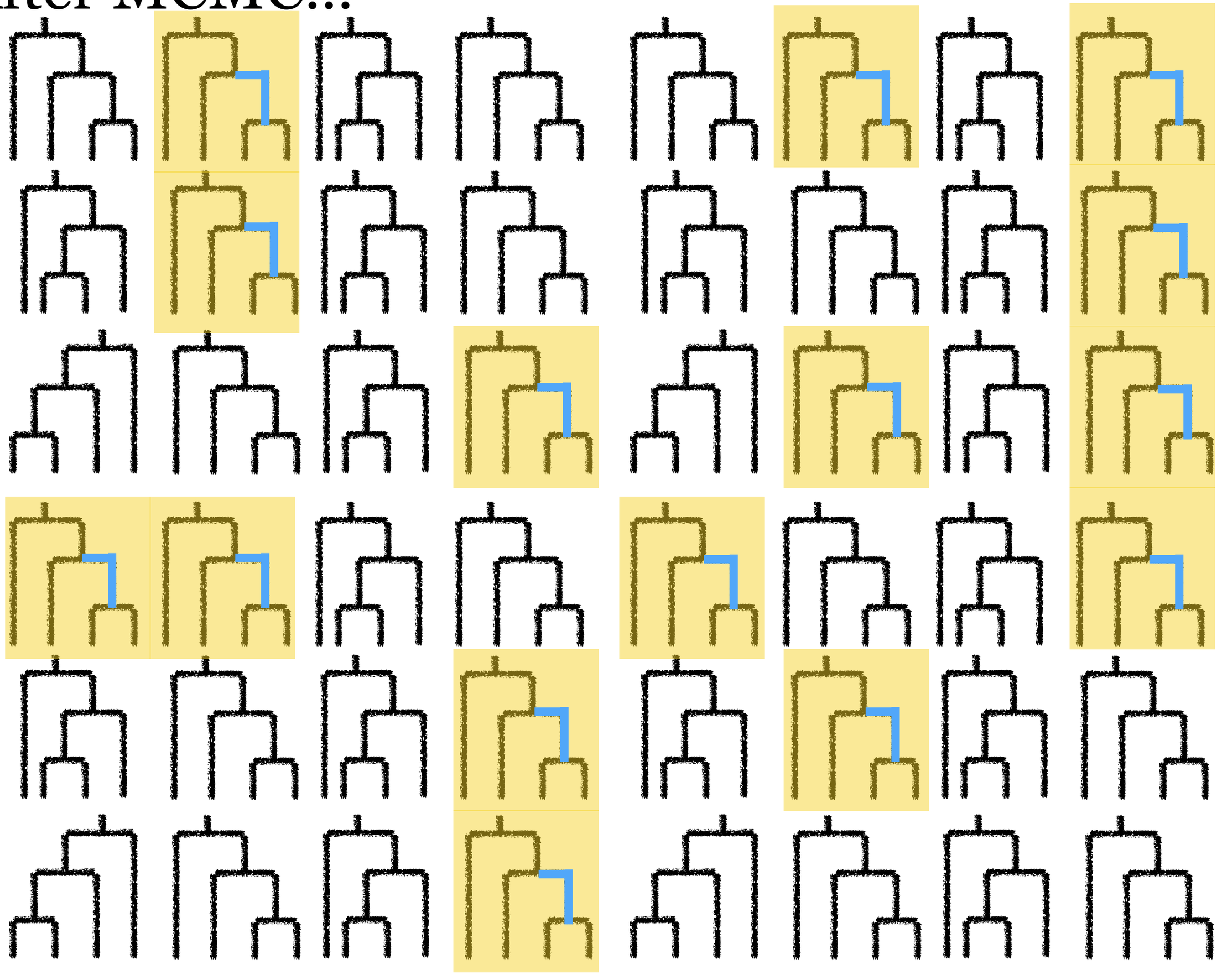
After MCMC…

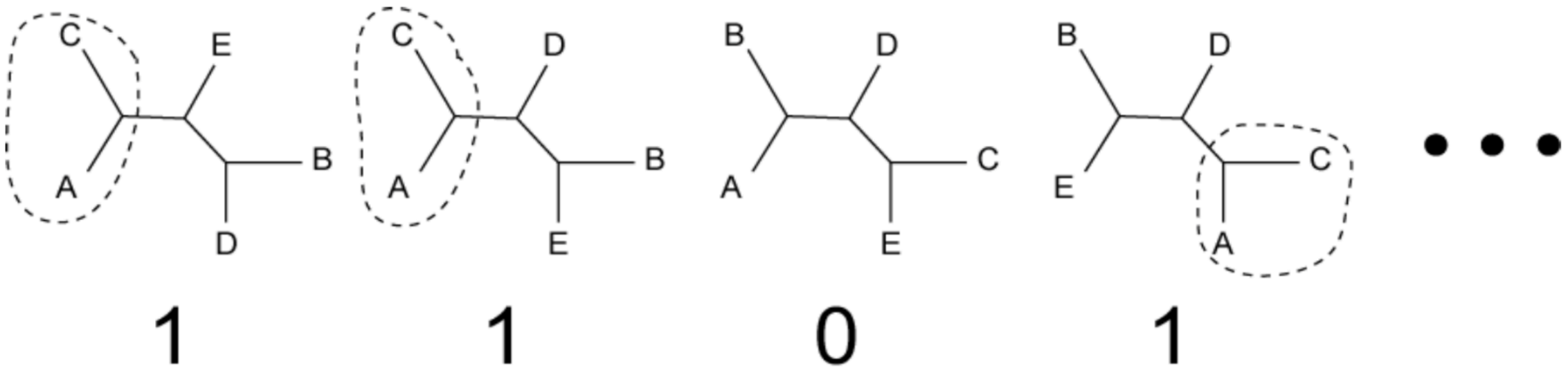


After MCMC…



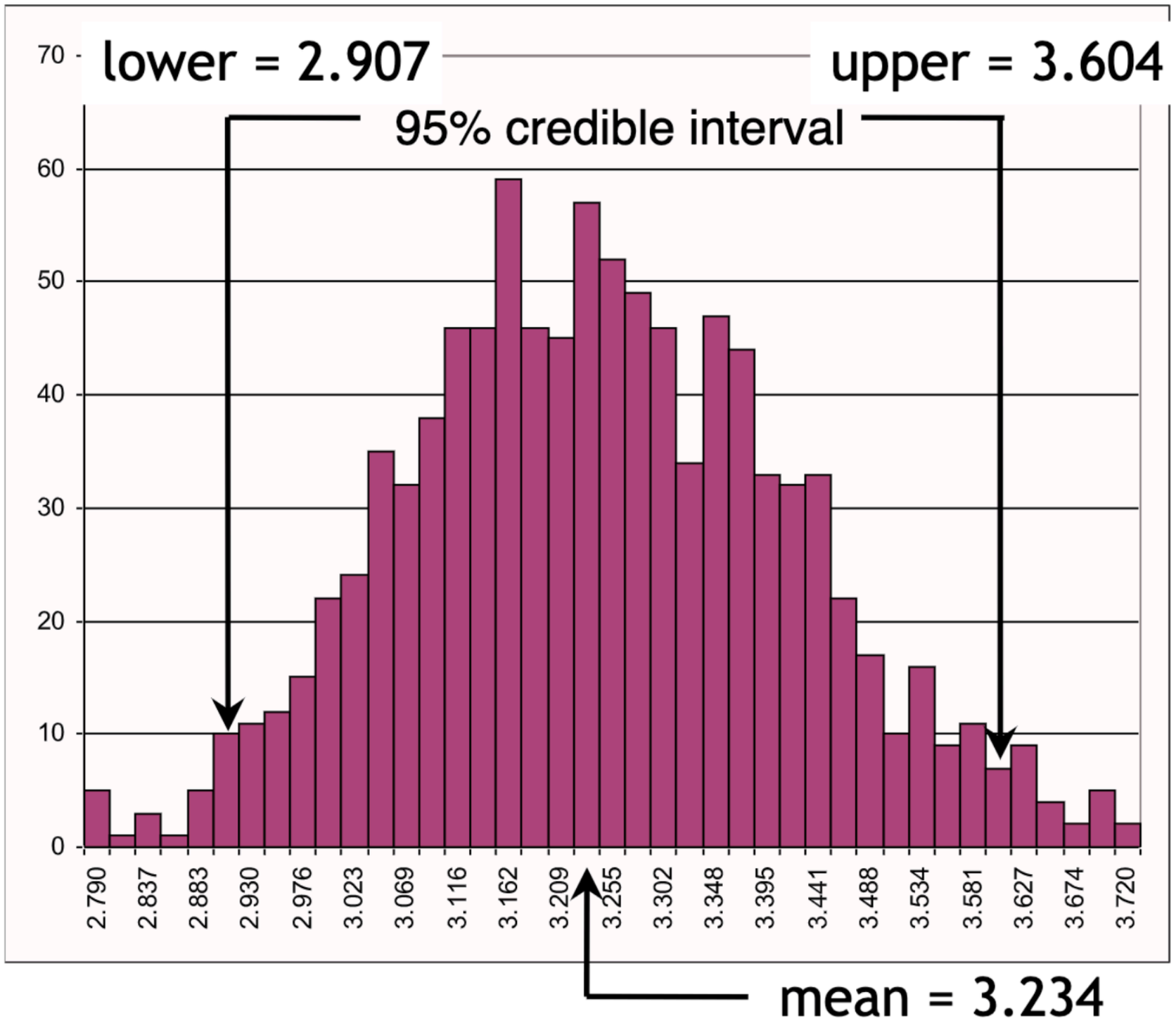
After MCMC…





The posterior probability of the split AC|BDE may be approximated by the fraction of trees sampled from the posterior that contain that split.

Marginal Posterior Distribution of κ



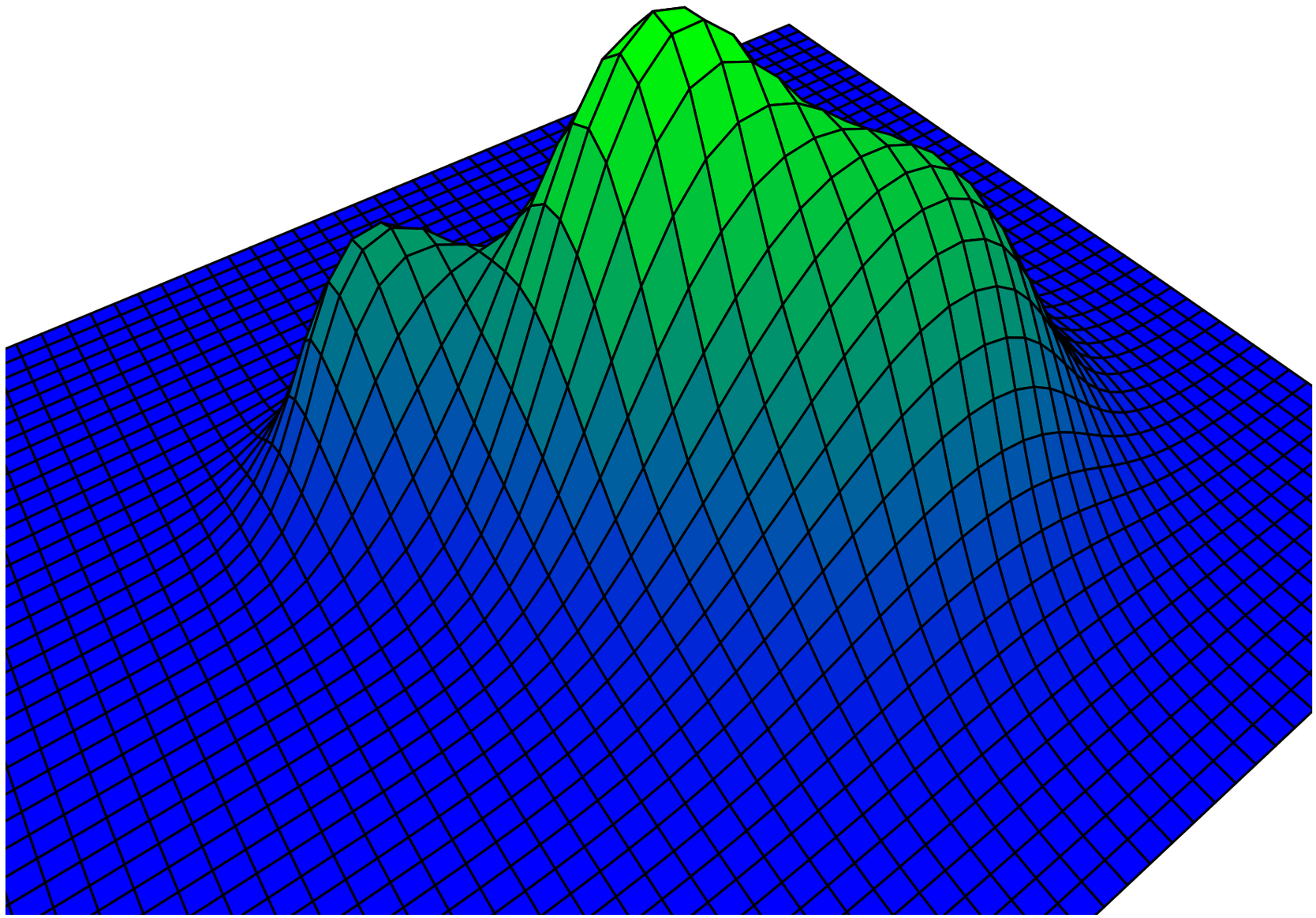
Histogram created from a sample of 1000 kappa values.

Data from Lewis, L., and Flechtner, V. 2002. Taxon 51: 443-451.

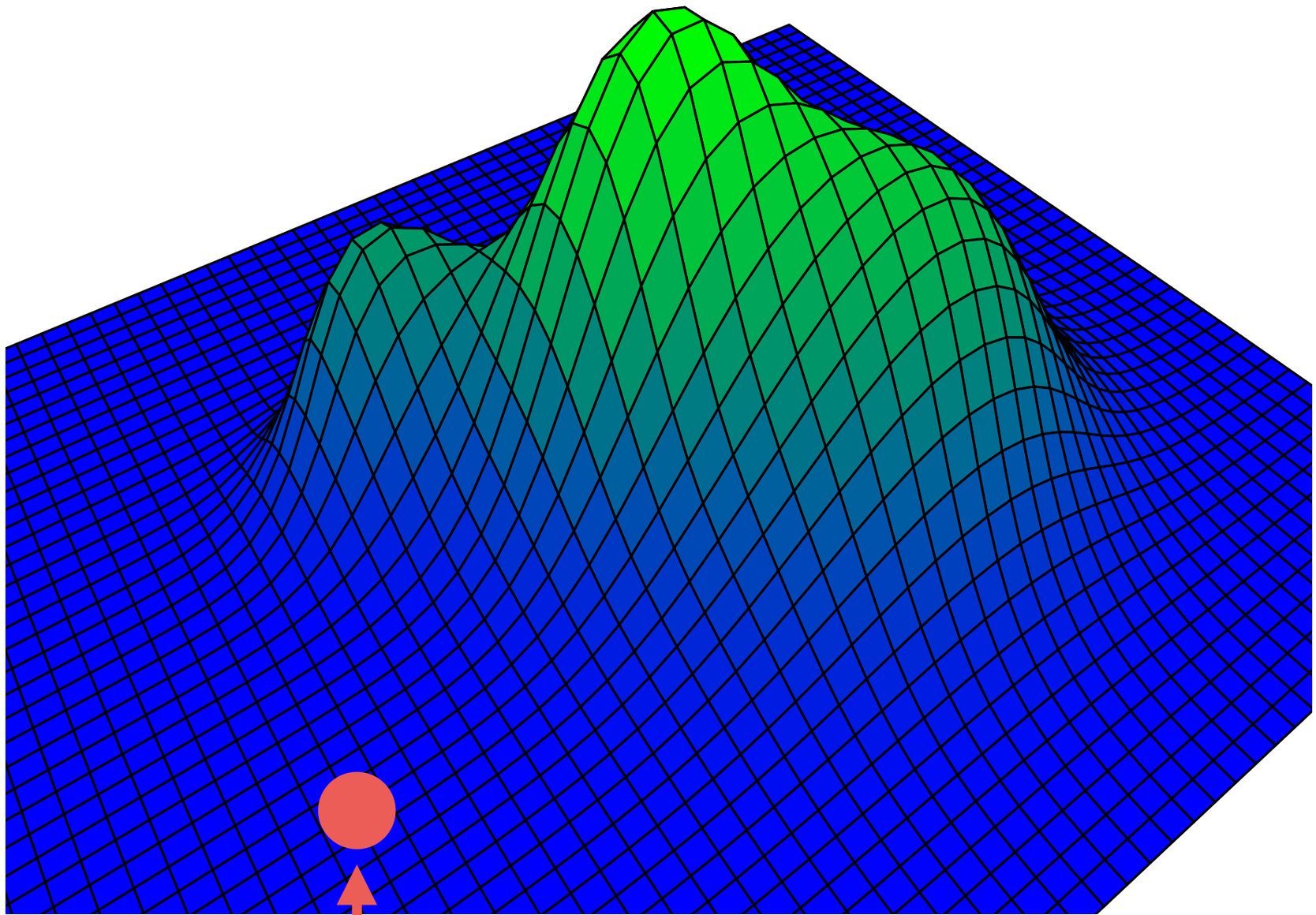
Why is MCMC so slow?

Why is MCMC so slow? Traverse tree space

Why is MCMC so slow? Traverse tree space

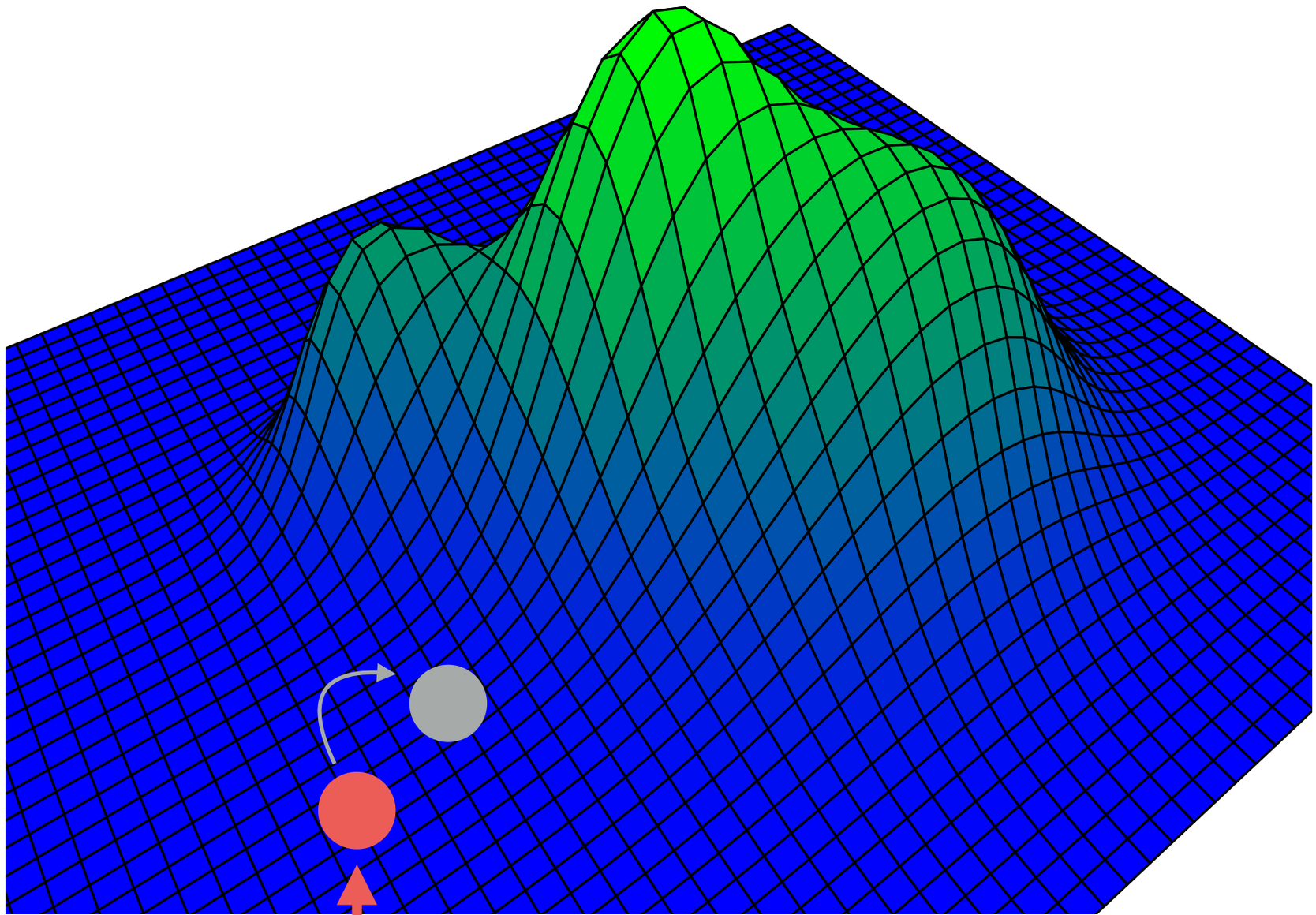


Why is MCMC so slow? Traverse tree space



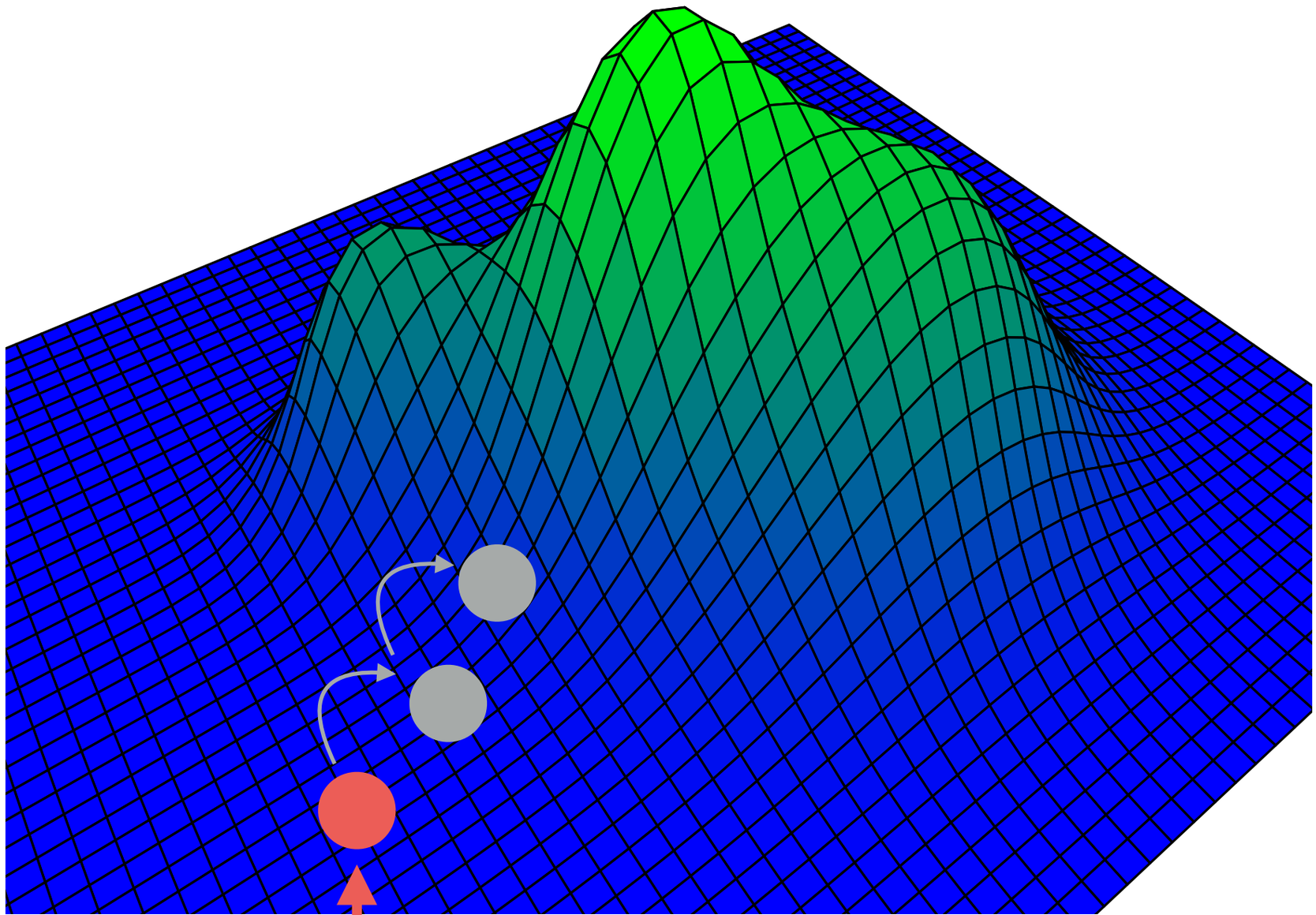
Starting tree

Why is MCMC so slow? Traverse tree space



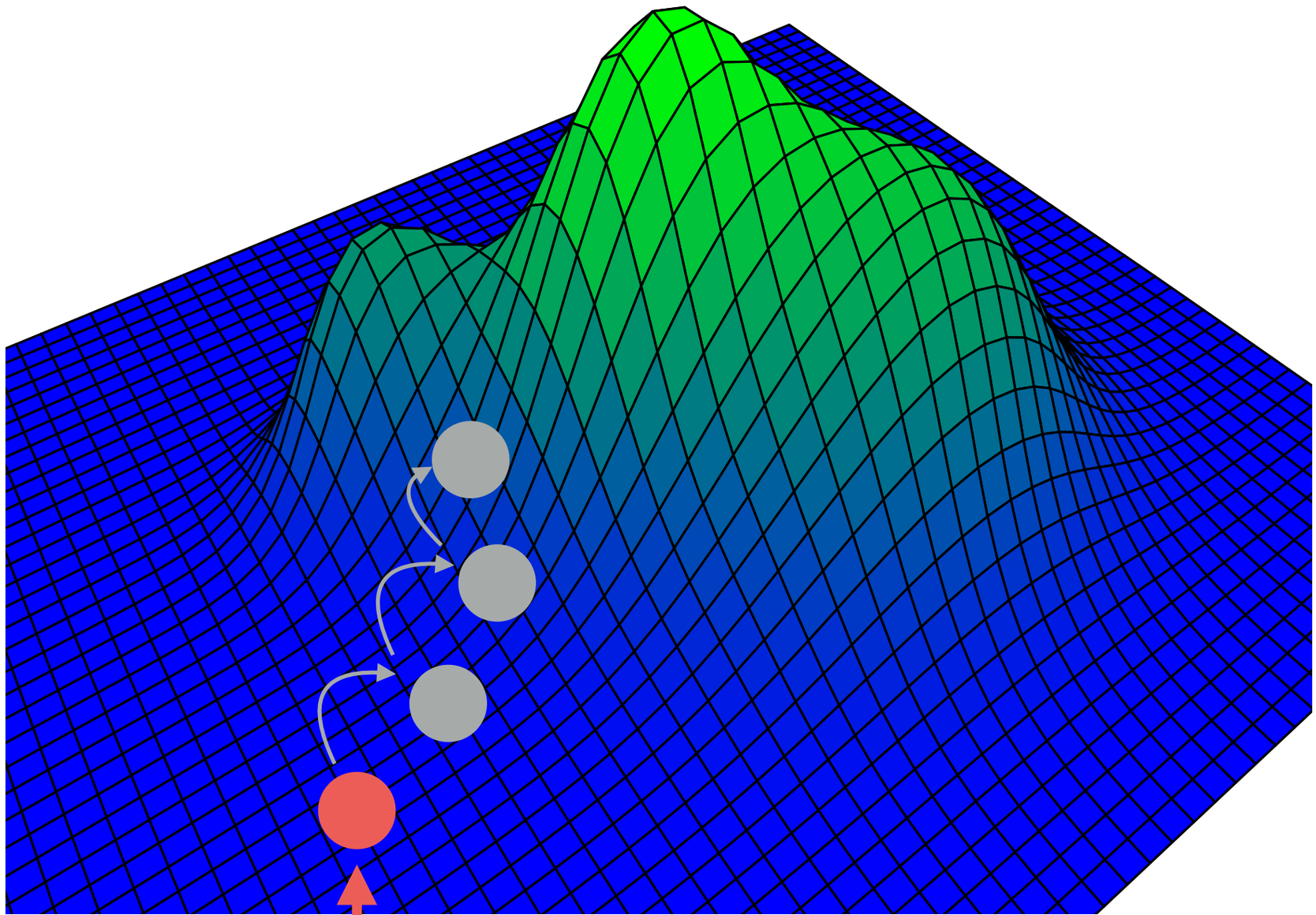
Starting tree

Why is MCMC so slow? Traverse tree space



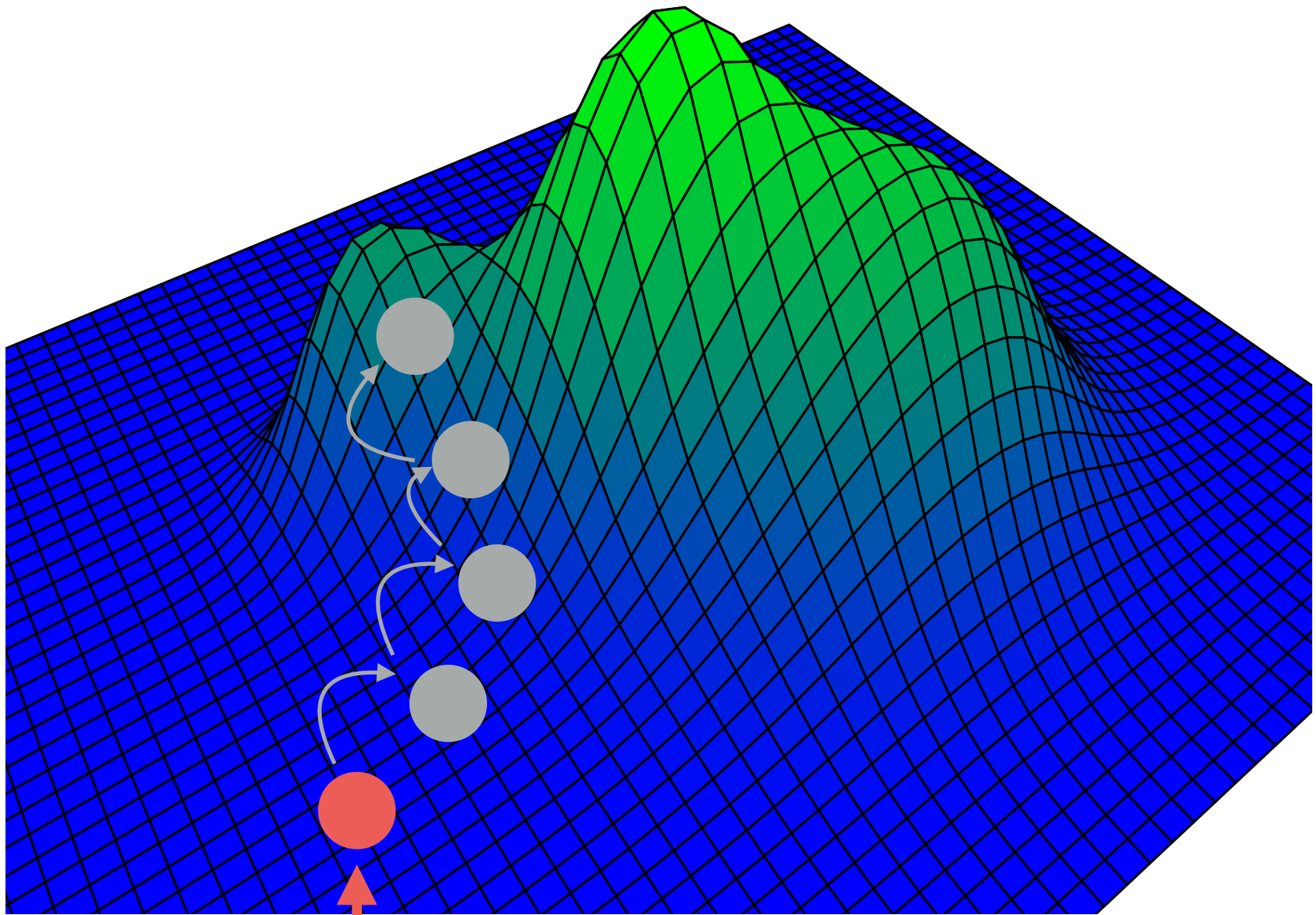
Starting tree

Why is MCMC so slow? Traverse tree space



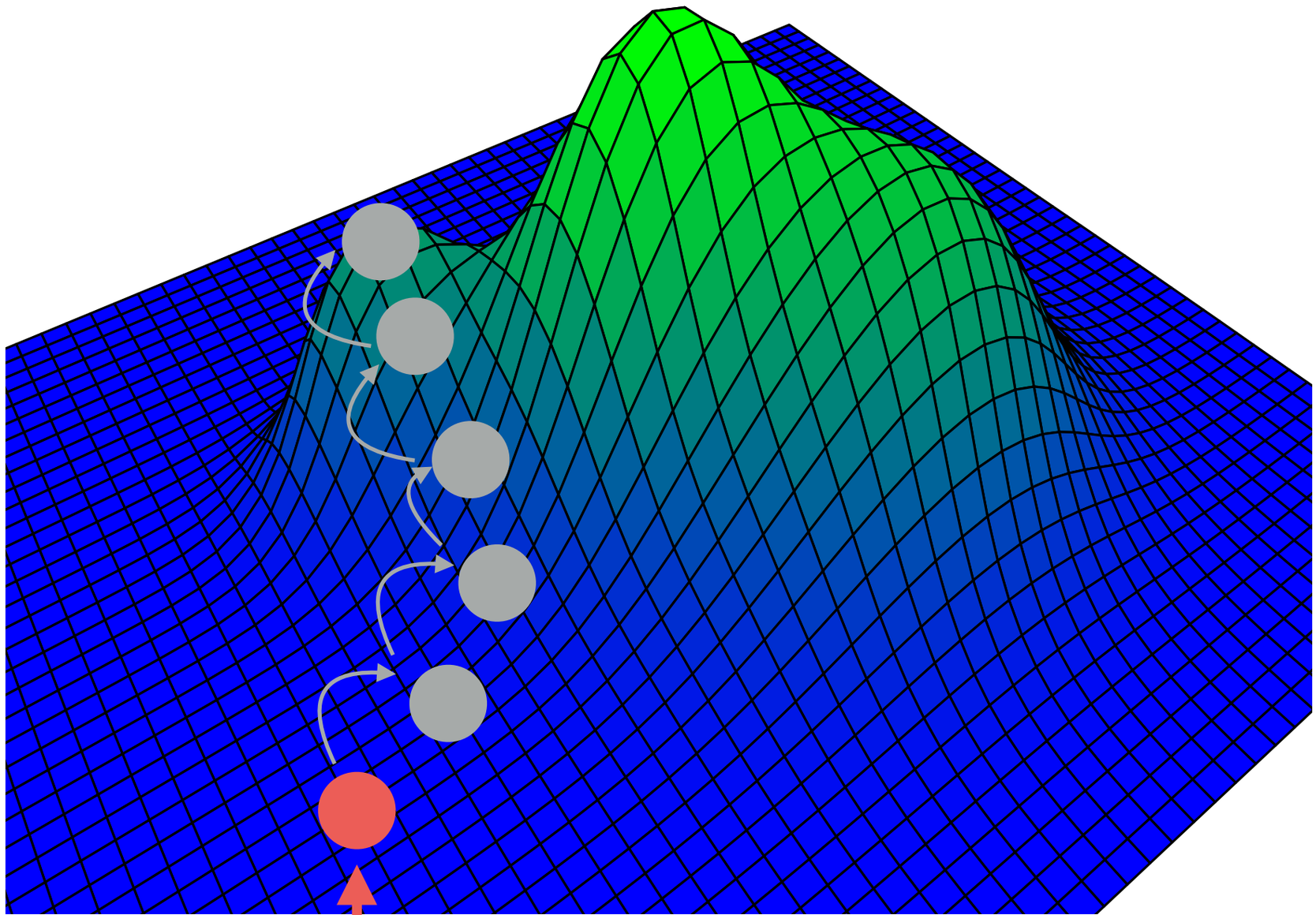
Starting tree

Why is MCMC so slow? Traverse tree space



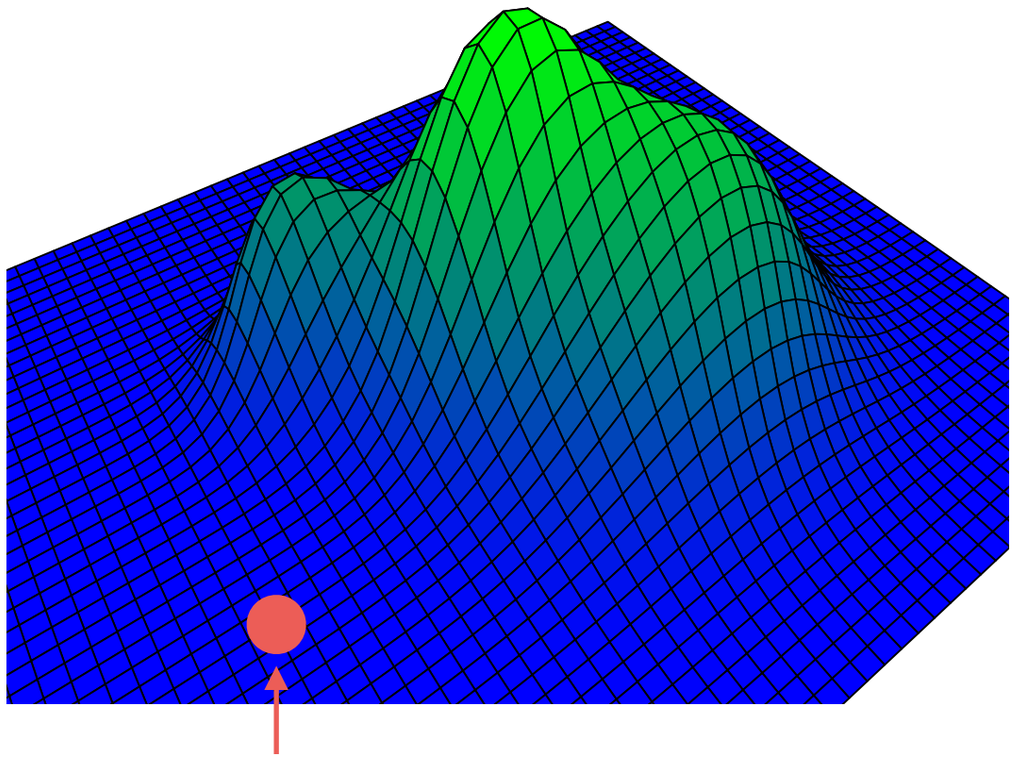
Starting tree

Why is MCMC so slow? Traverse tree space

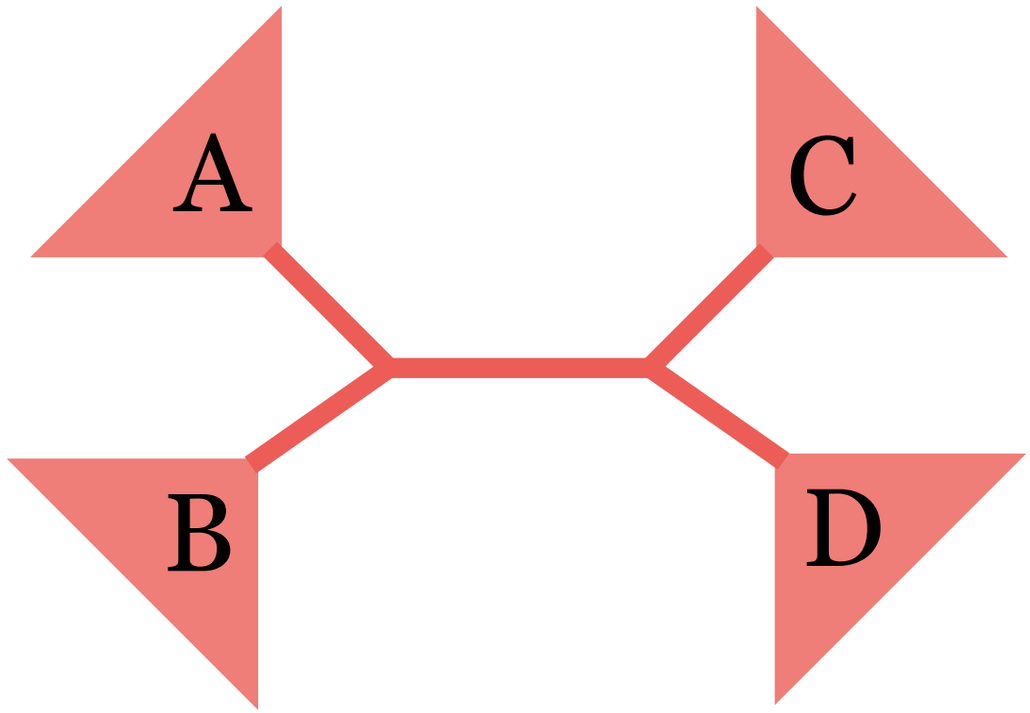


Starting tree

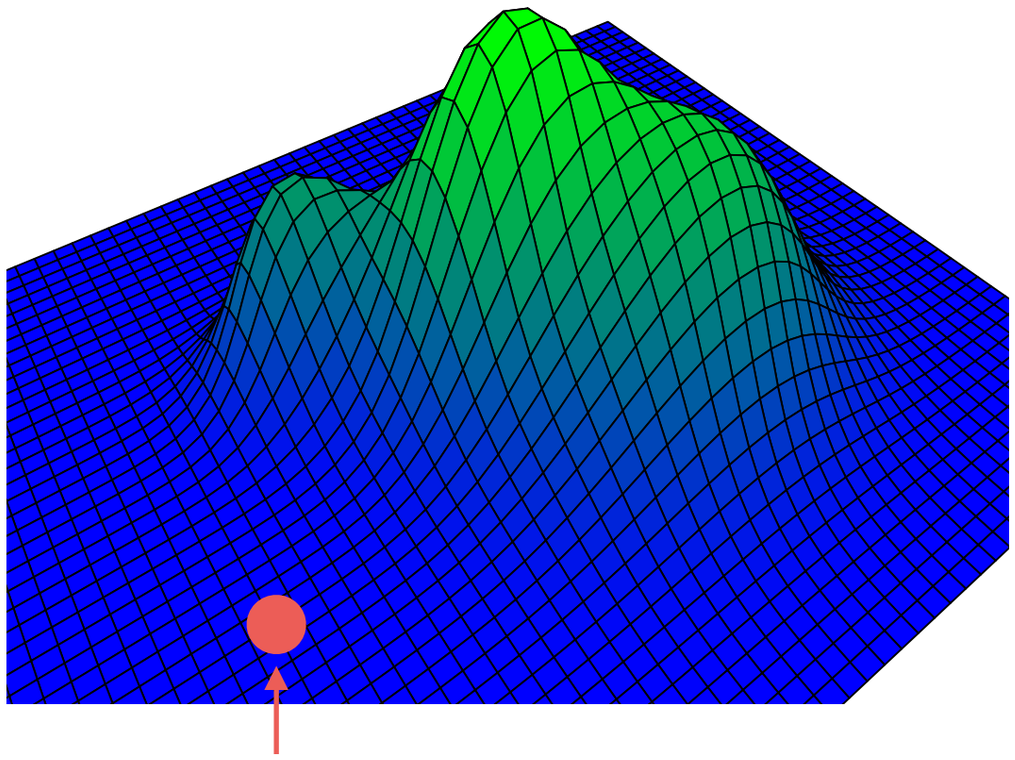
Why is MCMC so slow? Traverse tree space



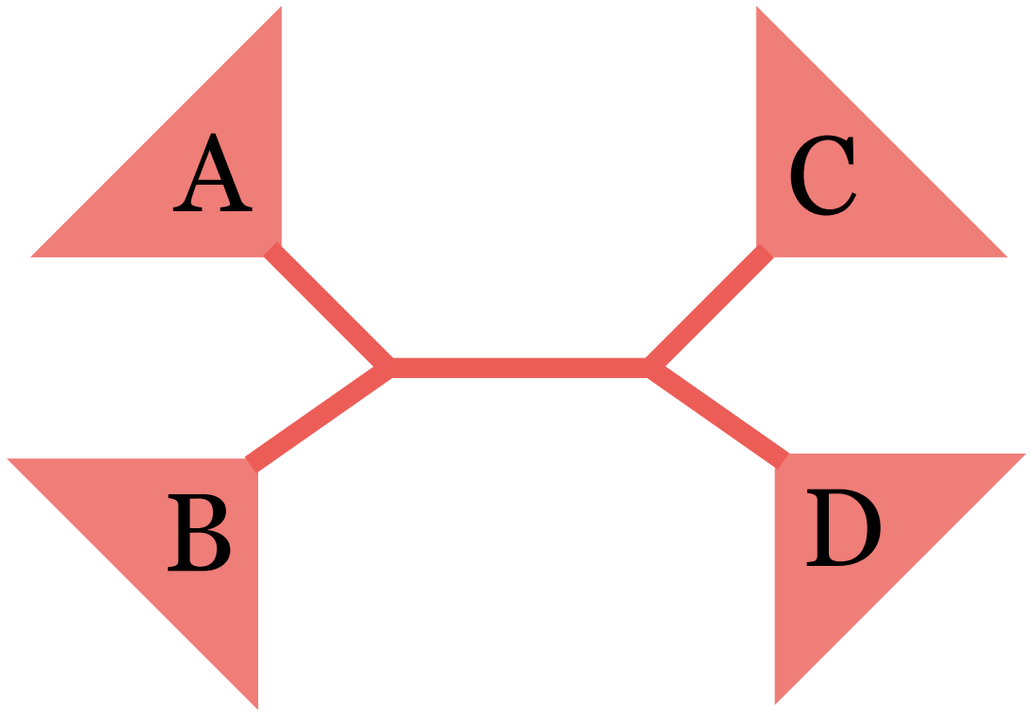
Starting tree



Why is MCMC so slow? Traverse tree space

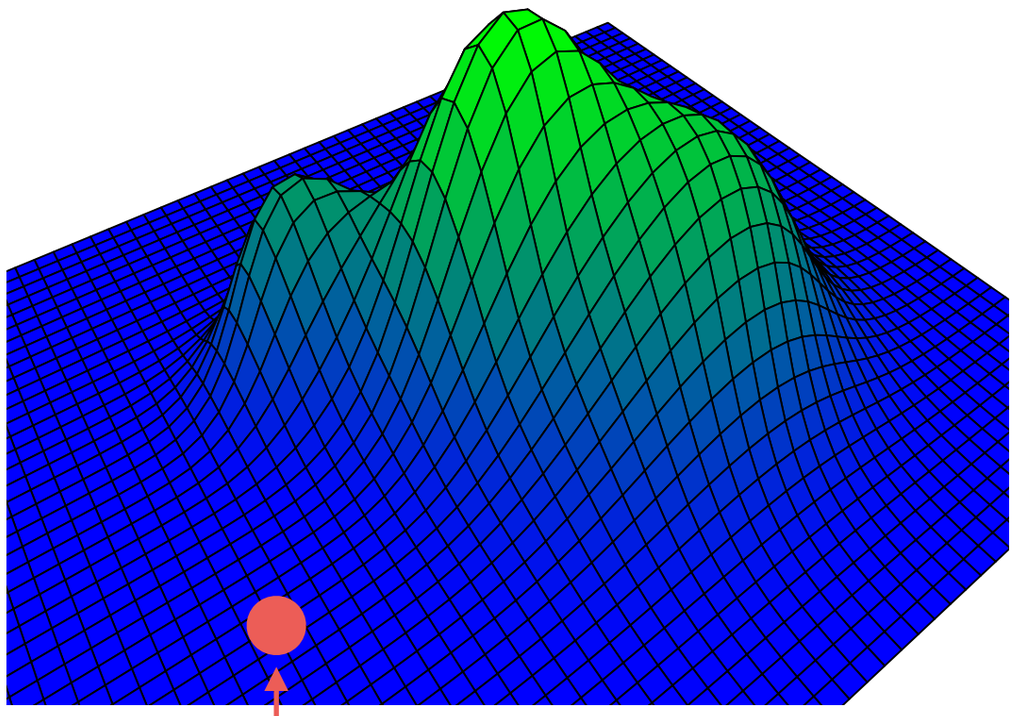


Starting tree

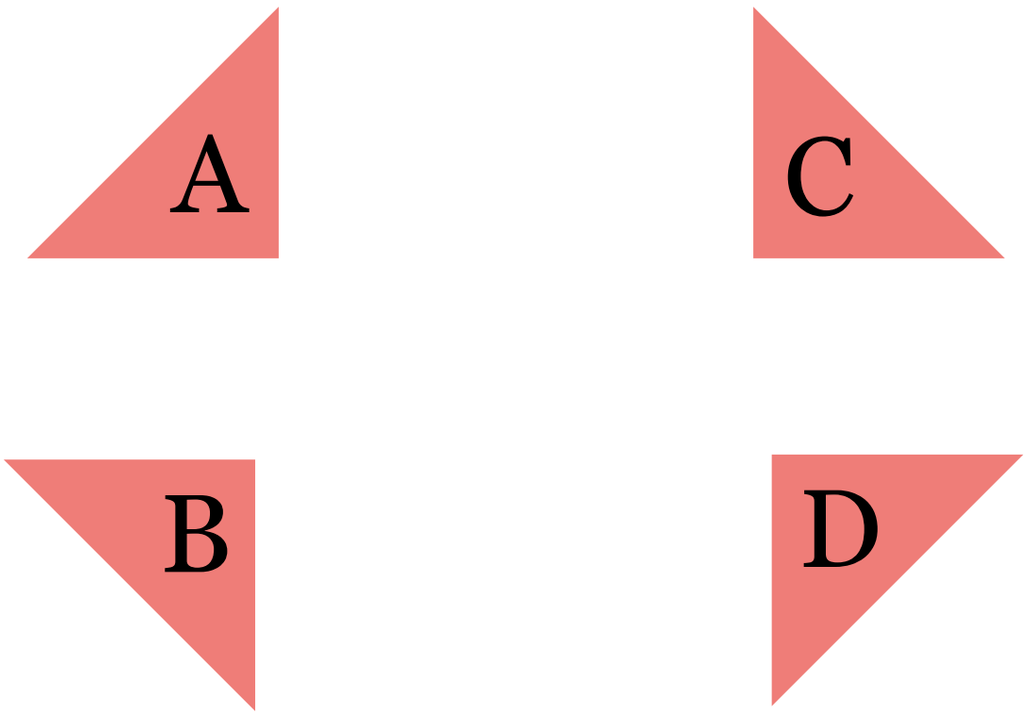


Nearest Neighbor Interchange

Why is MCMC so slow? Traverse tree space



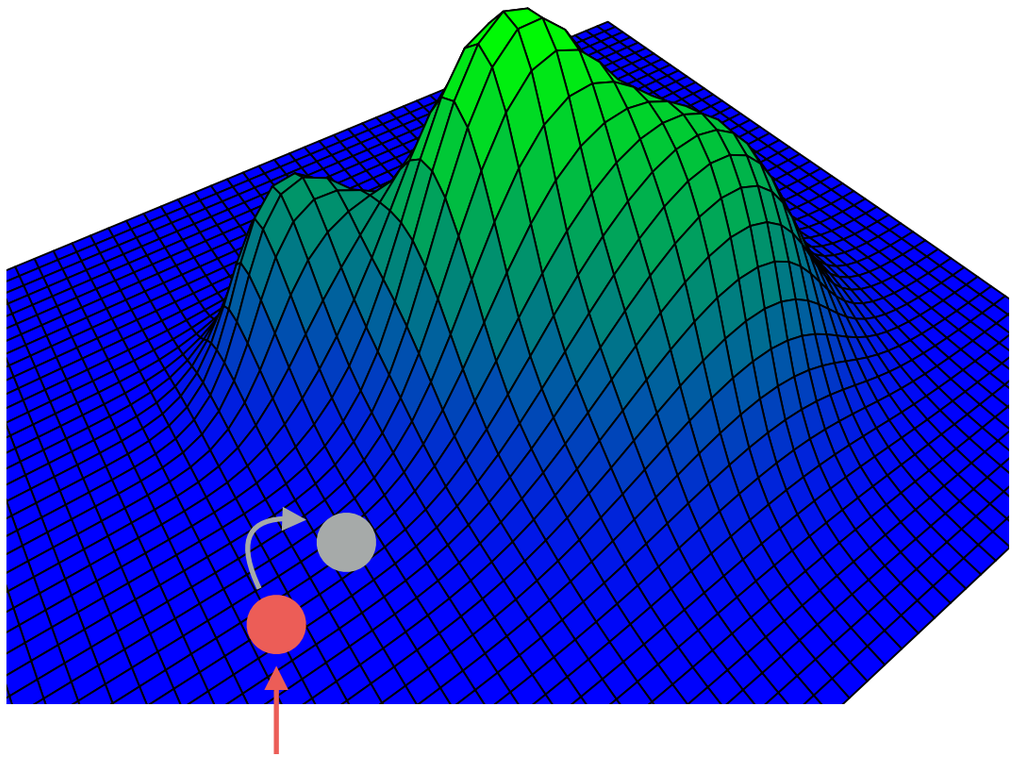
Starting tree



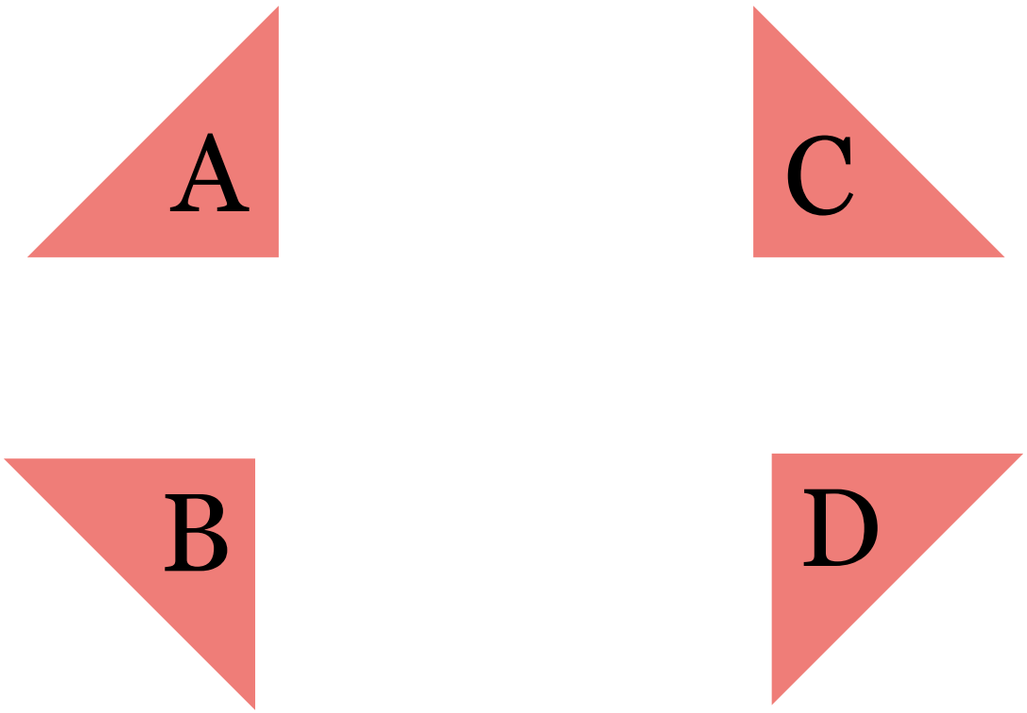
Nearest Neighbor Interchange

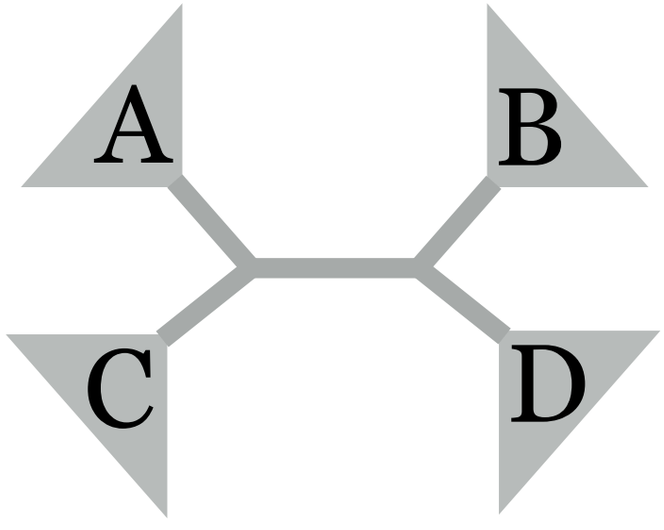
Why is MCMC so slow?

Traverse tree space



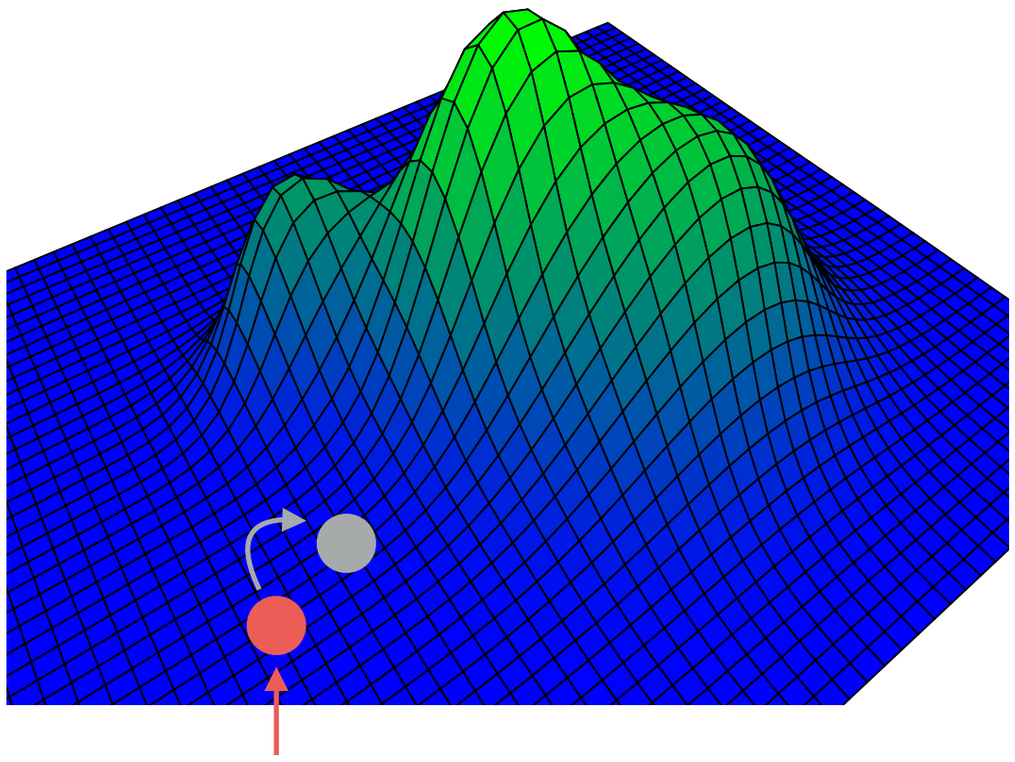
Starting tree



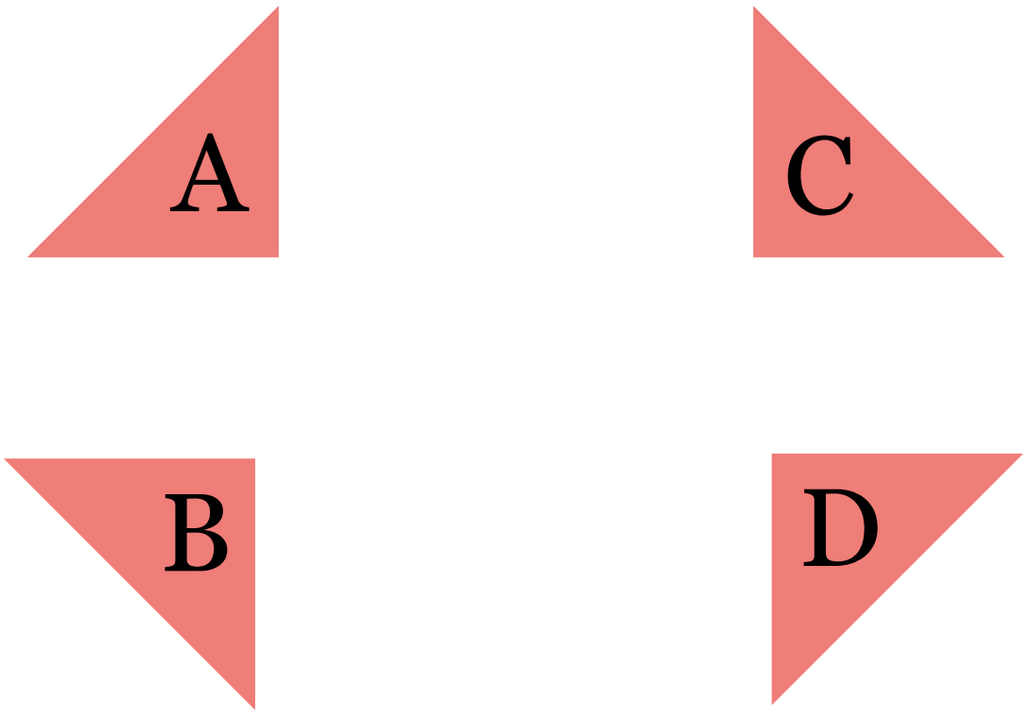


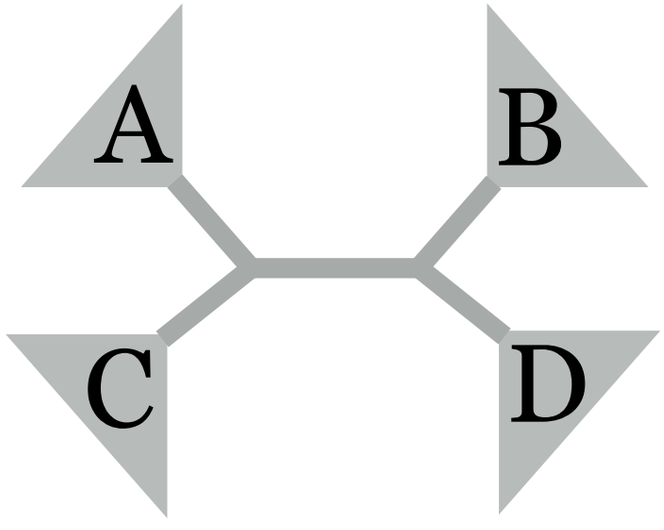
Nearest Neighbor Interchange

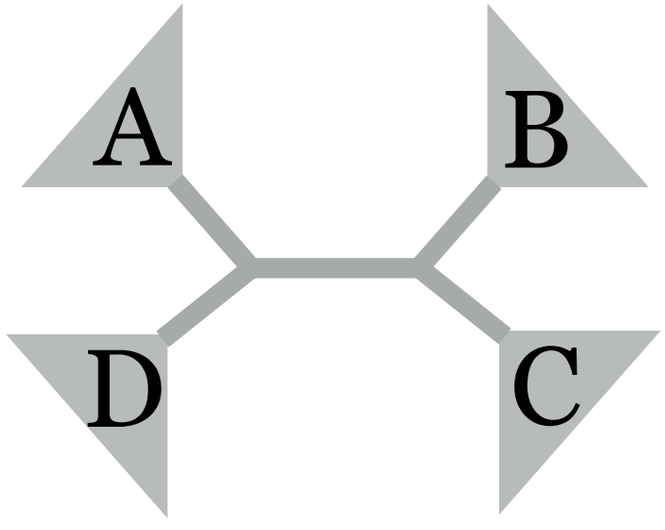
Why is MCMC so slow? Traverse tree space



Starting tree







Nearest Neighbor Interchange

Why is MCMC so slow?

Why is MCMC so slow? Tree space is huge

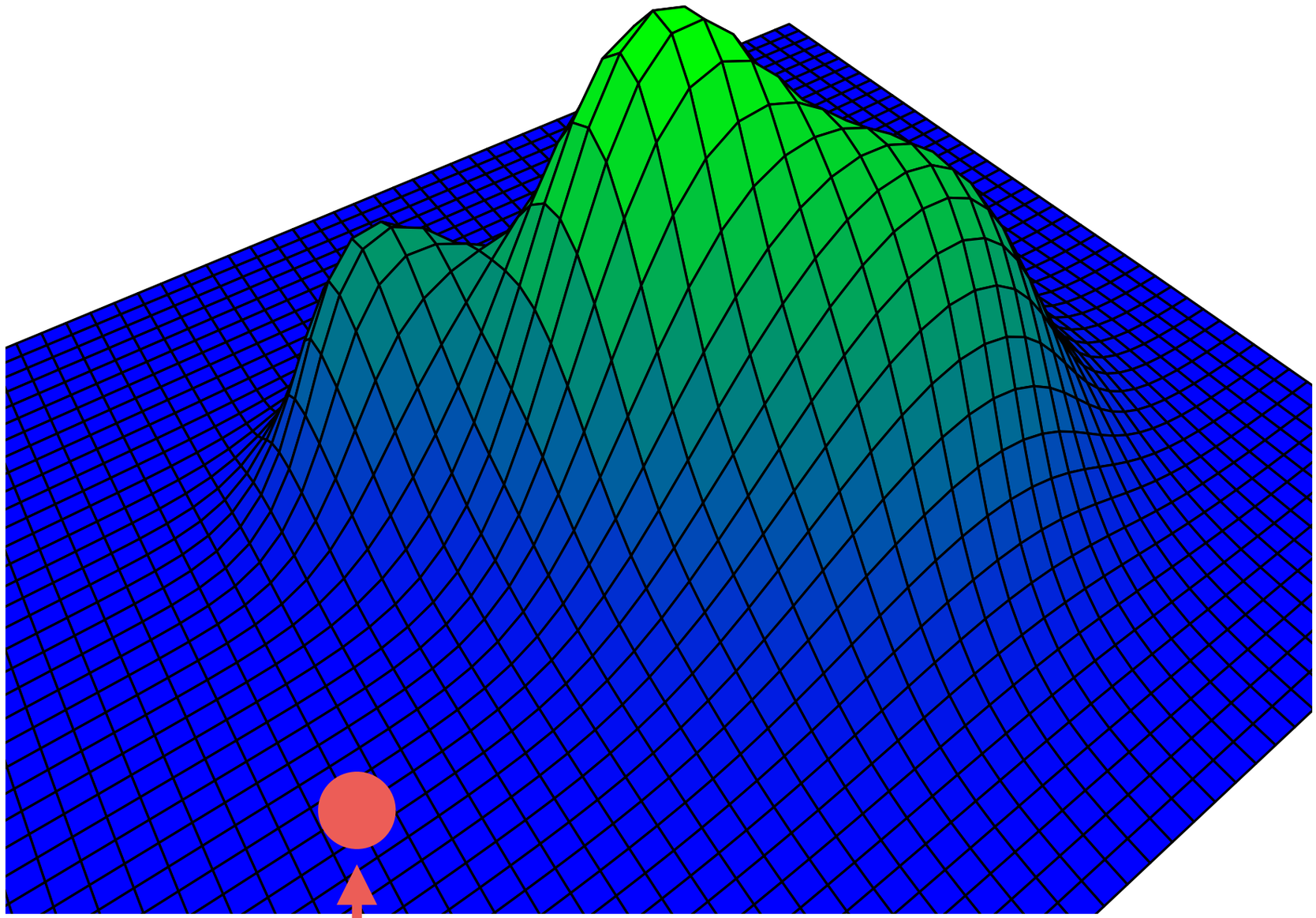
Why is MCMC so slow? Tree space is huge

# Species	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225
⋮	⋮	⋮
52	$>$ # atoms in universe	

Why is MCMC so slow?

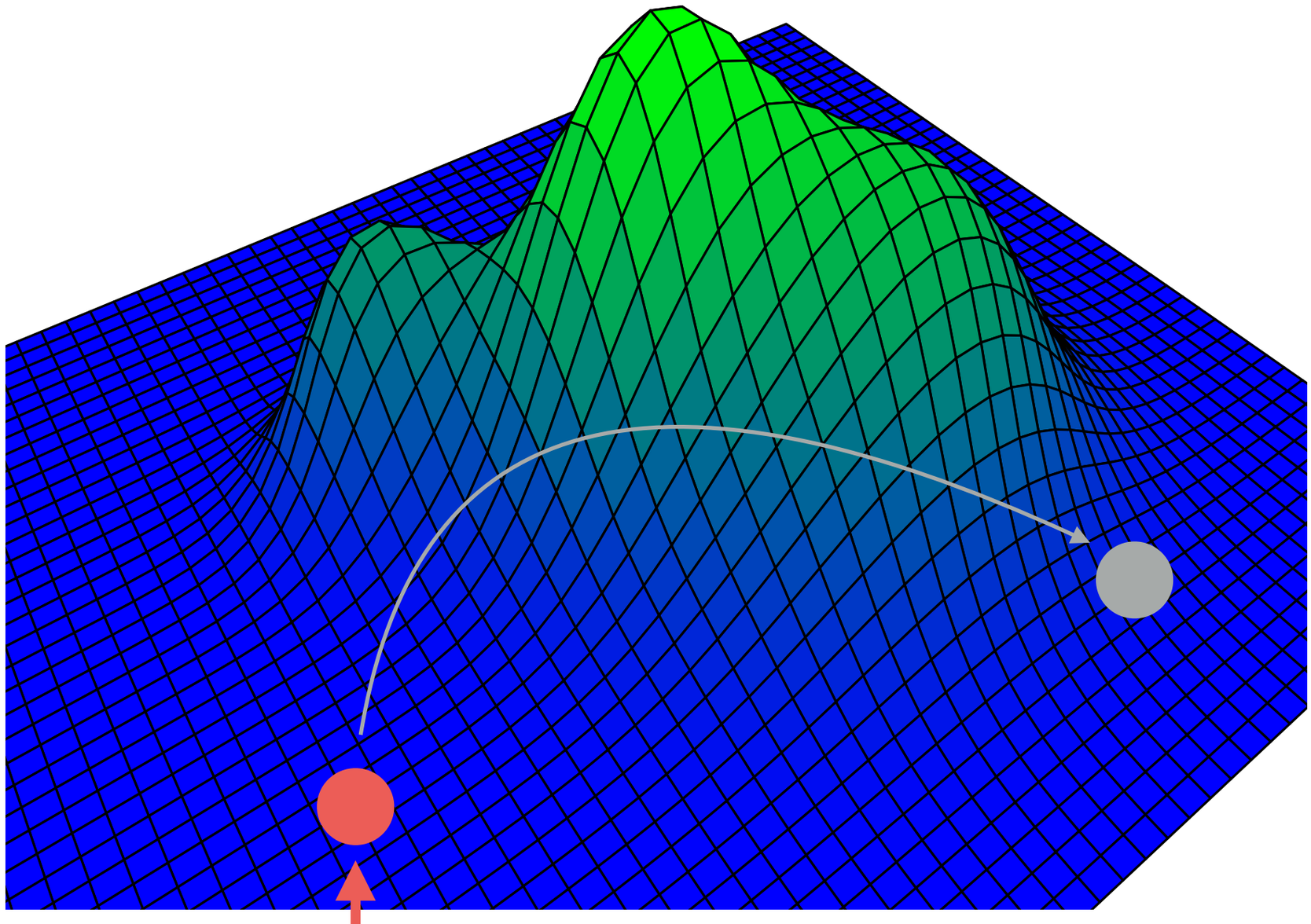
Why is MCMC so slow? Low acceptance of moves

Why is MCMC so slow? Low acceptance of moves



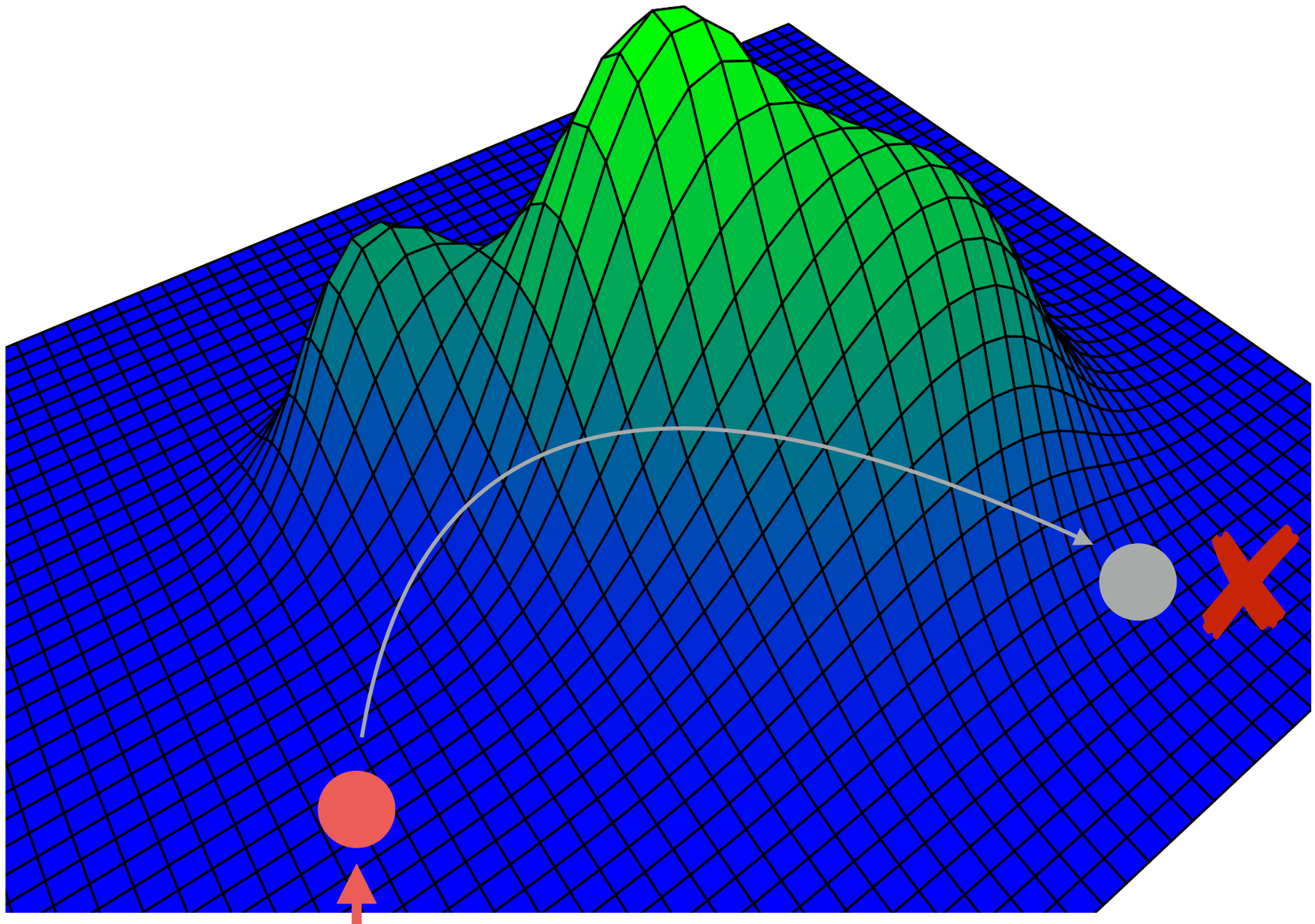
Starting tree

Why is MCMC so slow? Low acceptance of moves



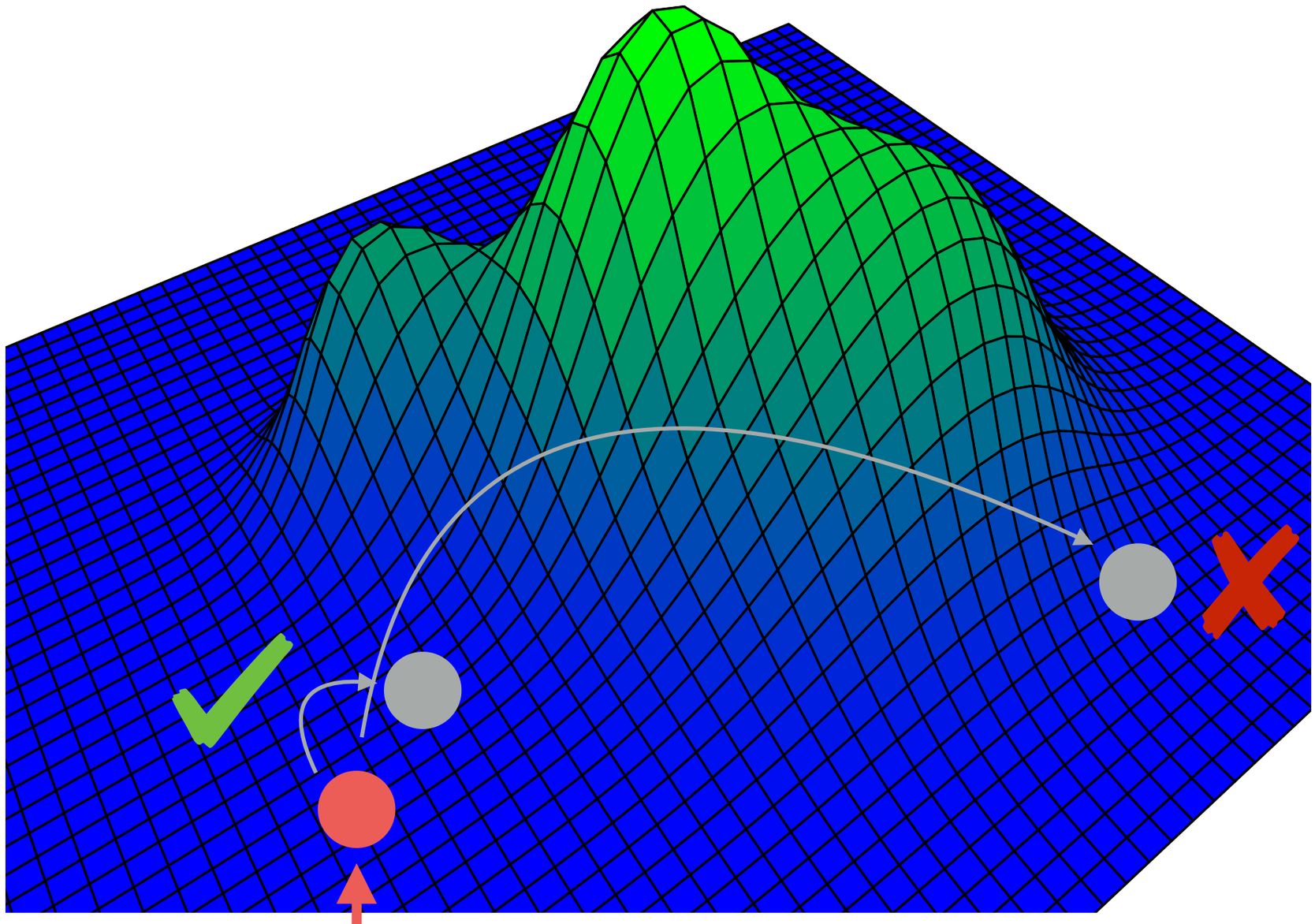
Starting tree

Why is MCMC so slow? Low acceptance of moves



Starting tree

Why is MCMC so slow? Low acceptance of moves



Starting tree

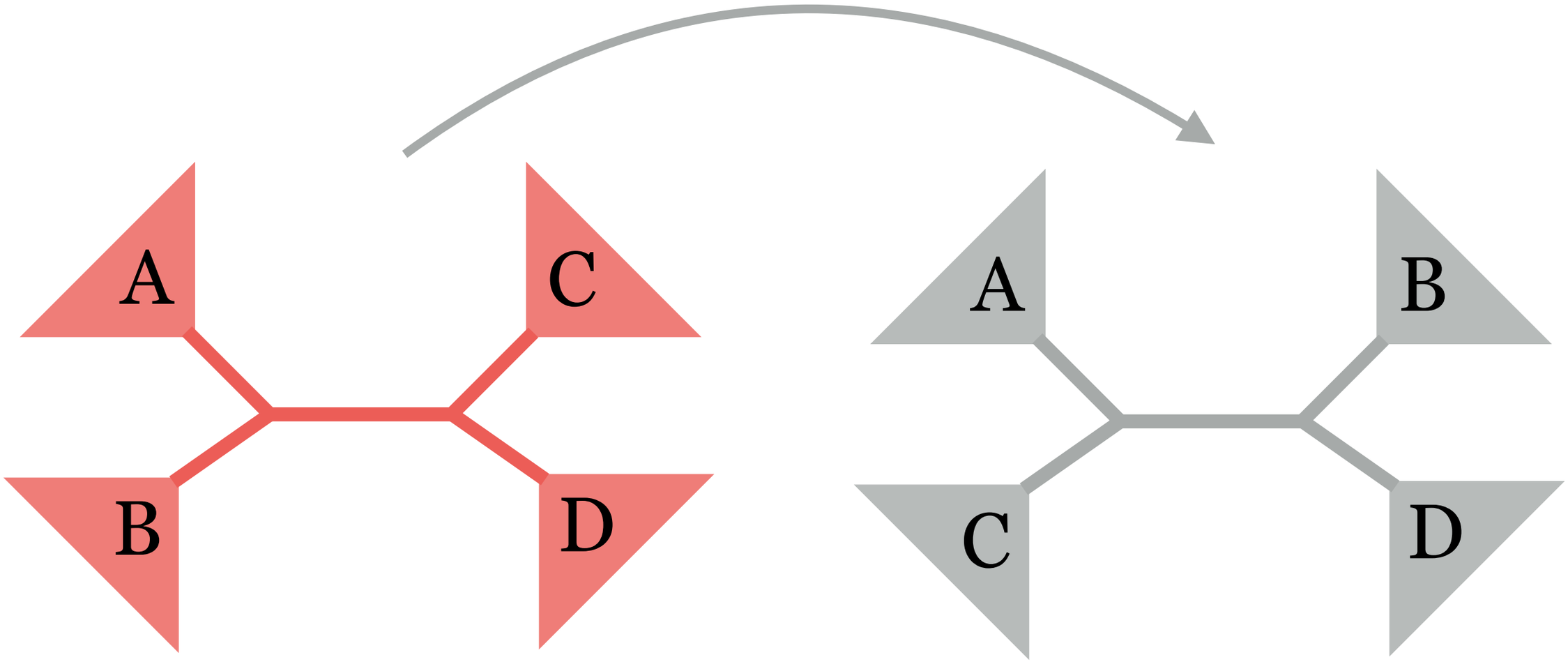
Why is MCMC so slow?

Why is MCMC so slow?

Small neighborhood implies very dependent sample

Why is MCMC so slow?

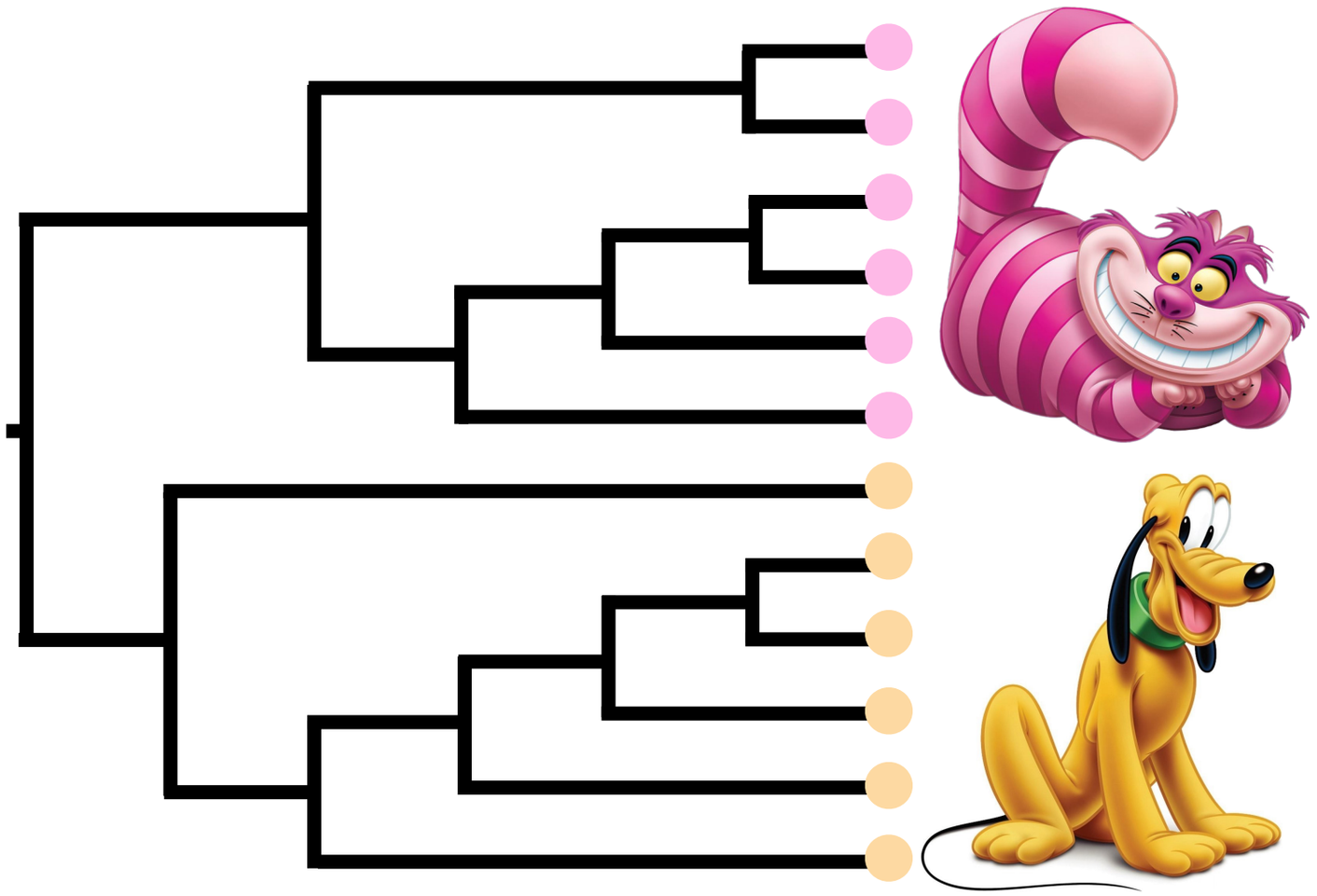
Small neighborhood implies very dependent sample



Why is MCMC so slow?

1) Huge tree space size
2) Low acceptance of moves unless small neighborhood
3) Small neighborhood implies very dependent sample, which means small effective sample size

We need a gigantic chain because the space is huge and we are making tiny moves



12 taxa *Carnivora*

MCMC efficiency ~0.025%

(250 from 1 million post-burnin generations)

ESS

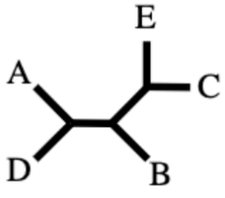
(Larget, 2013)

Priors

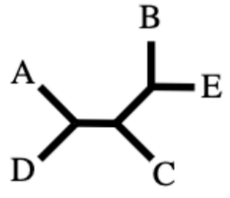
Common Priors

- **Discrete uniform** for topologies
  - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support [0,∞)
  - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and GTR relative rates

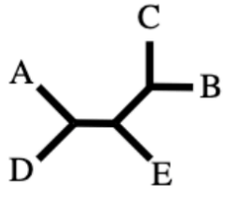
Discrete Uniform distribution for **topologies**



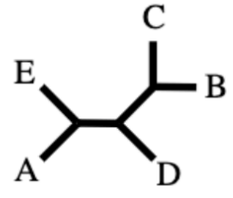
$$\frac{1}{15}$$



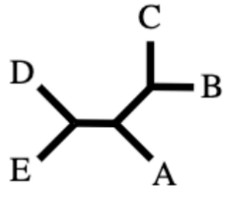
$$\frac{1}{15}$$



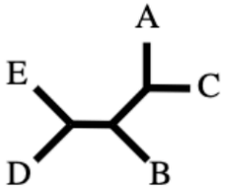
$$\frac{1}{15}$$



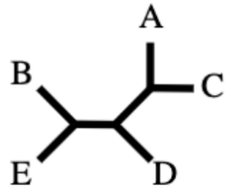
$$\frac{1}{15}$$



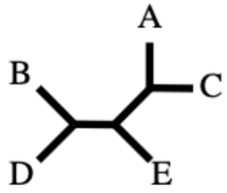
$$\frac{1}{15}$$



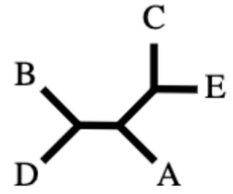
$$\frac{1}{15}$$



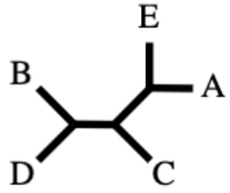
$$\frac{1}{15}$$



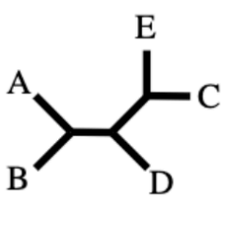
$$\frac{1}{15}$$



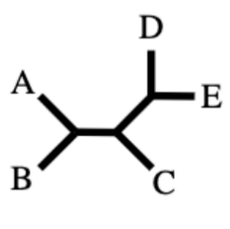
$$\frac{1}{15}$$



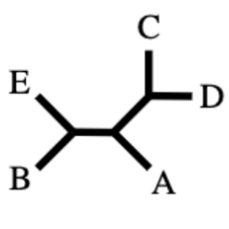
$$\frac{1}{15}$$



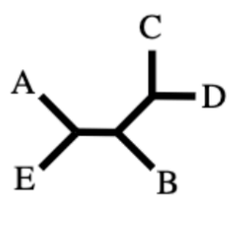
$$\frac{1}{15}$$



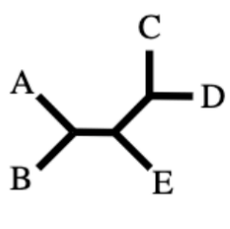
$$\frac{1}{15}$$



$$\frac{1}{15}$$

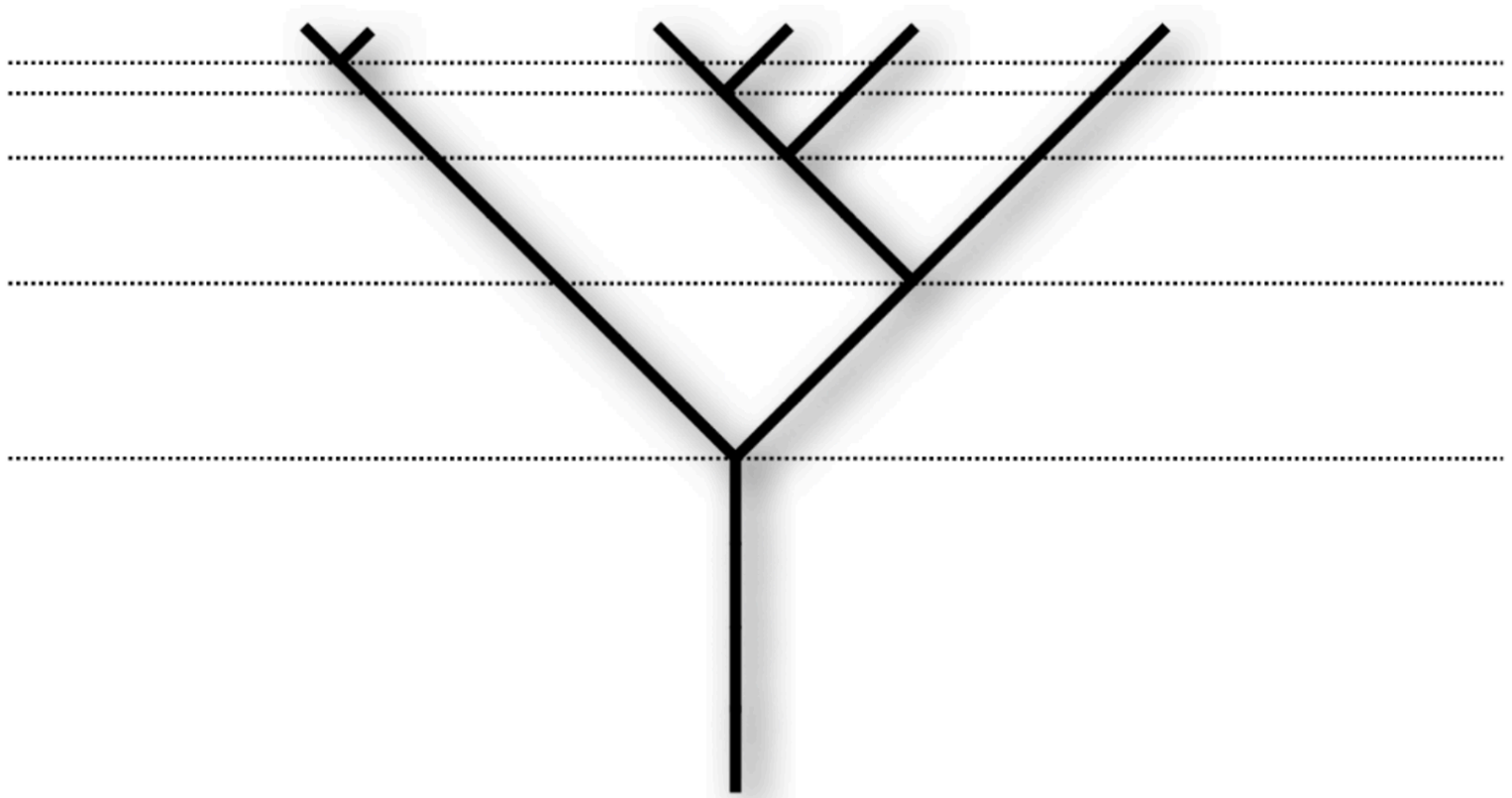


$$\frac{1}{15}$$



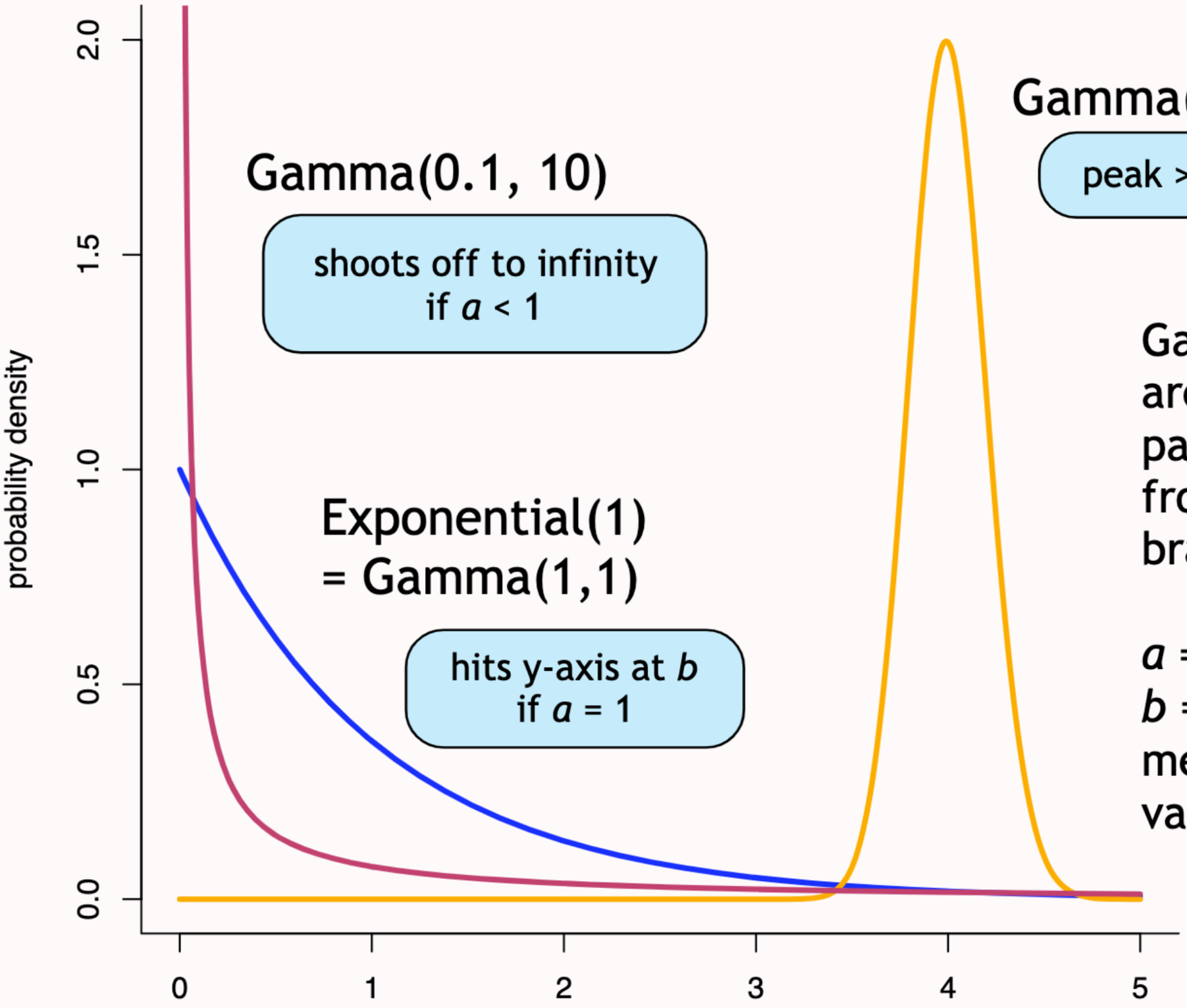
$$\frac{1}{15}$$

Yule model provides joint prior for both **topology** and **divergence times**



The rate of speciation under the Yule model (λ) is constant and applies equally and independently to each lineage. Thus, speciation events get closer together in time as the tree grows because more lineages are available to speciate.

Gamma(a,b) distributions



Gamma distributions are ideal for parameters that range from 0 to infinity (e.g. branch lengths)

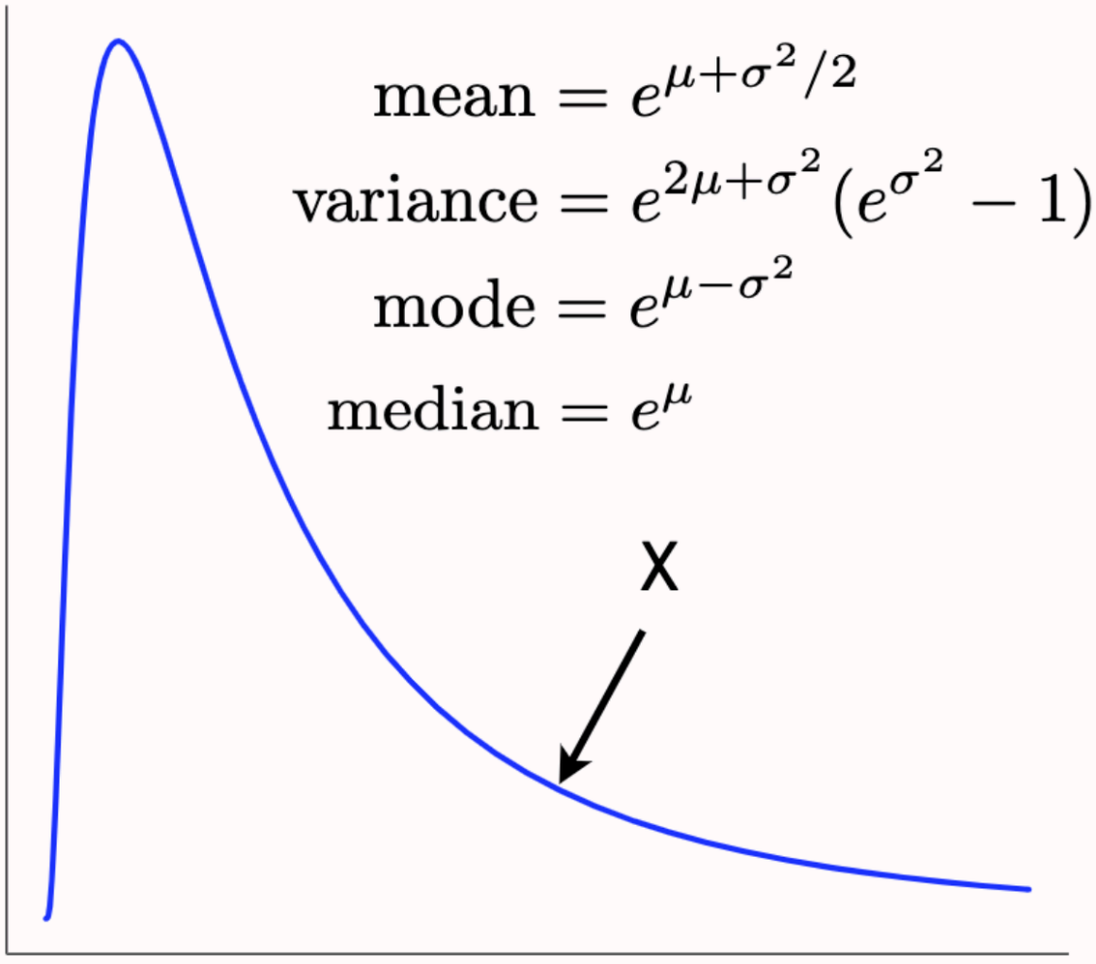
a = shape
b = scale
mean* = ab
variance* = ab^2

*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2, respectively.

Log-normal distribution

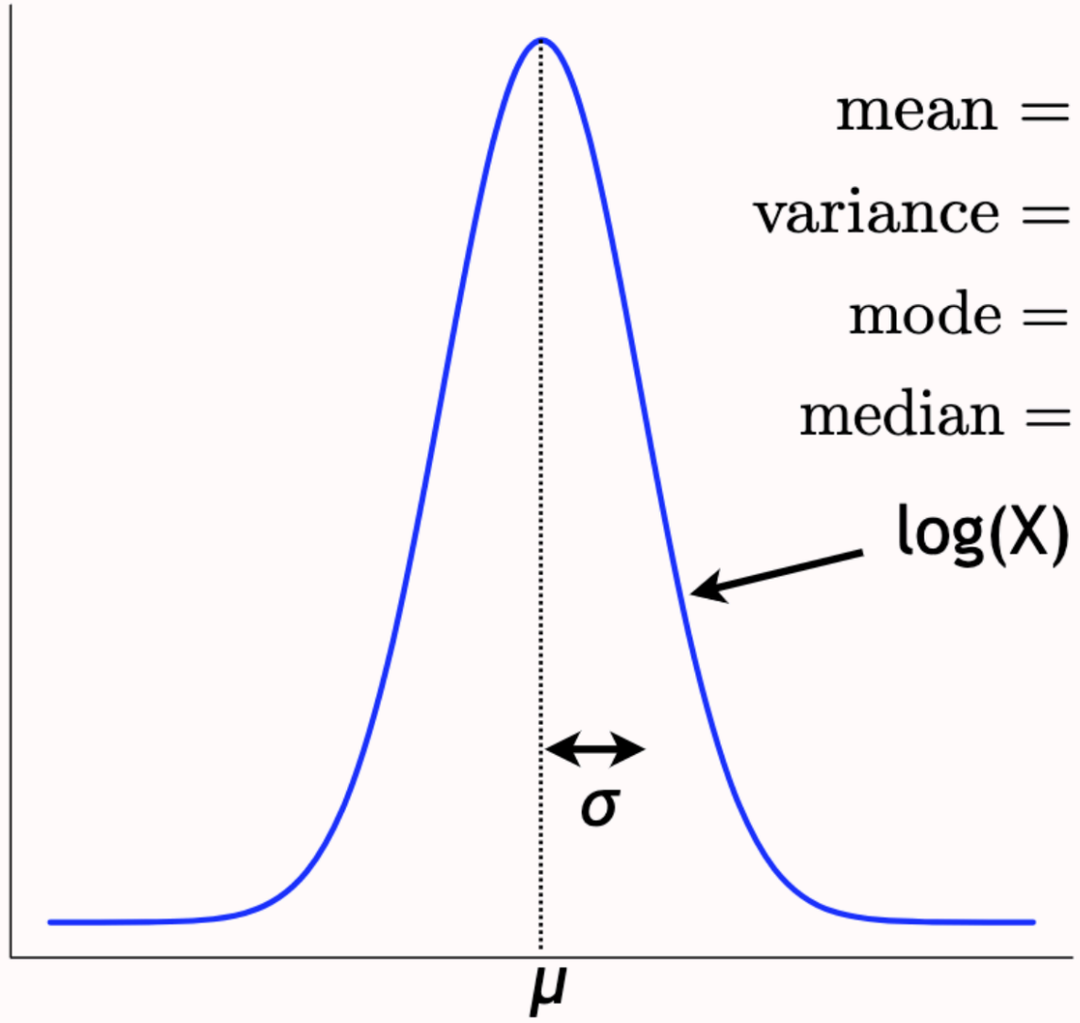
If **X** is **log-normal** with *parameters* μ and σ...

$$\text{mean} = e^{\mu+\sigma^2/2}$$
$$\text{variance} = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$
$$\text{mode} = e^{\mu-\sigma^2}$$
$$\text{median} = e^{\mu}$$



...then **log(X)** is **normal** with *mean* μ and *standard deviation* σ.

$$\text{mean} = \mu$$
$$\text{variance} = \sigma^2$$
$$\text{mode} = \mu$$
$$\text{median} = \mu$$

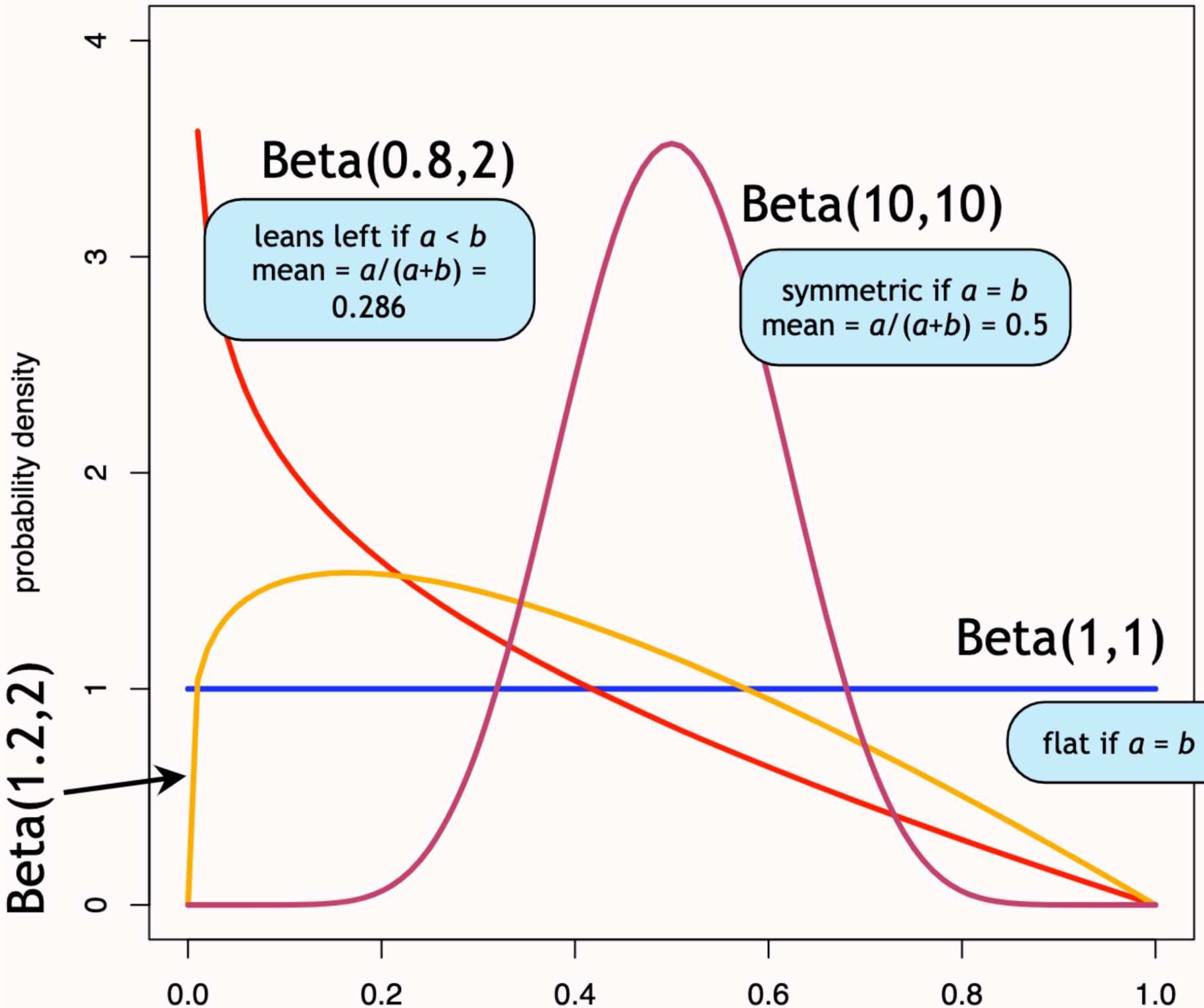


Important: *μ and σ do **not** represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

To choose μ and σ to yield a particular mean (m) and variance (v) for X, use these formulas:

$$\mu = \log(m^2) - \log(m) - \frac{\log(v+m^2) - \log(m^2)}{2}$$
$$\sigma^2 = \log(v+m^2) - \log(m^2)$$

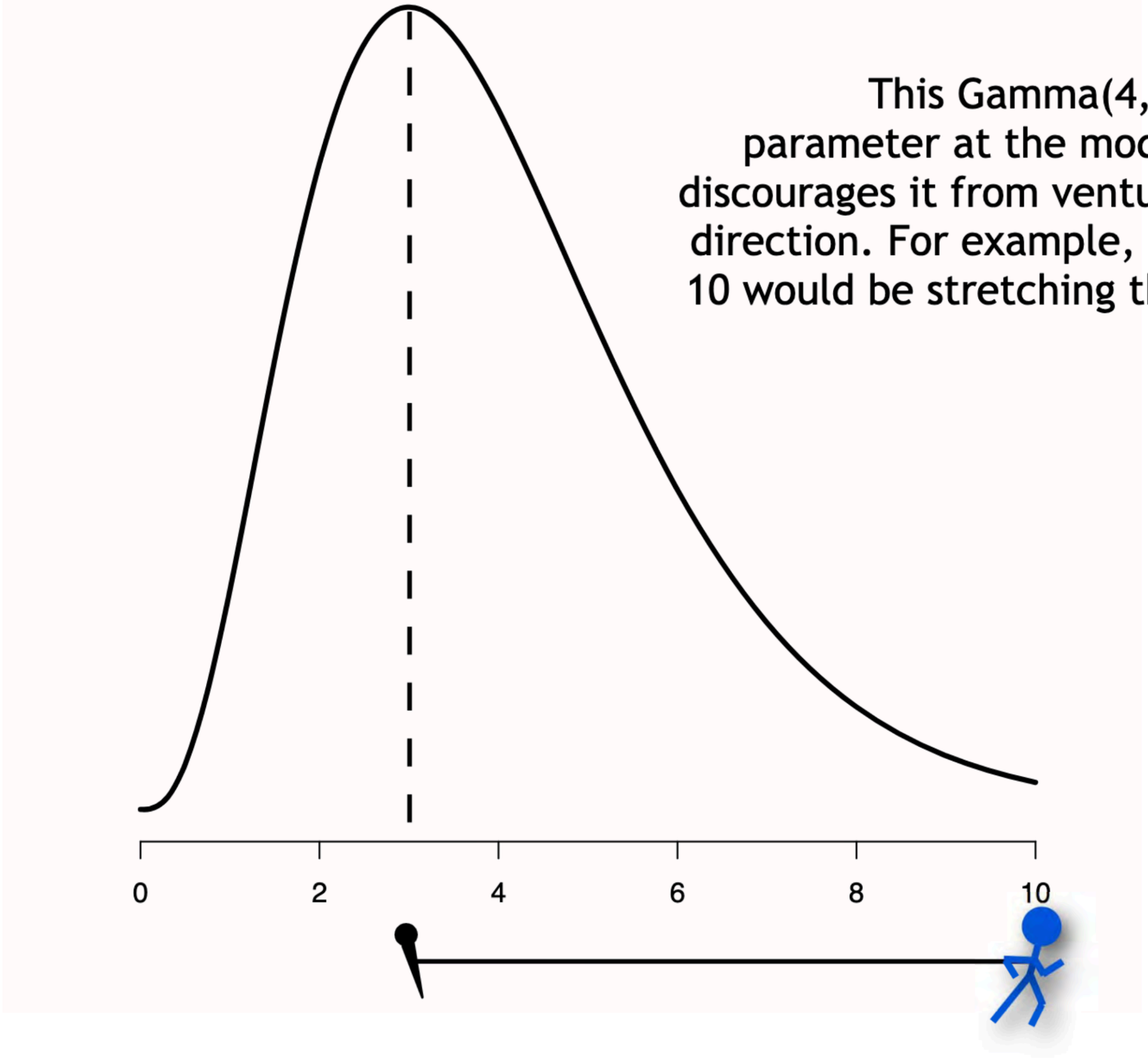
Beta(a,b) gallery



Beta distributions are appropriate for proportions, which are constrained to the interval [0,1].

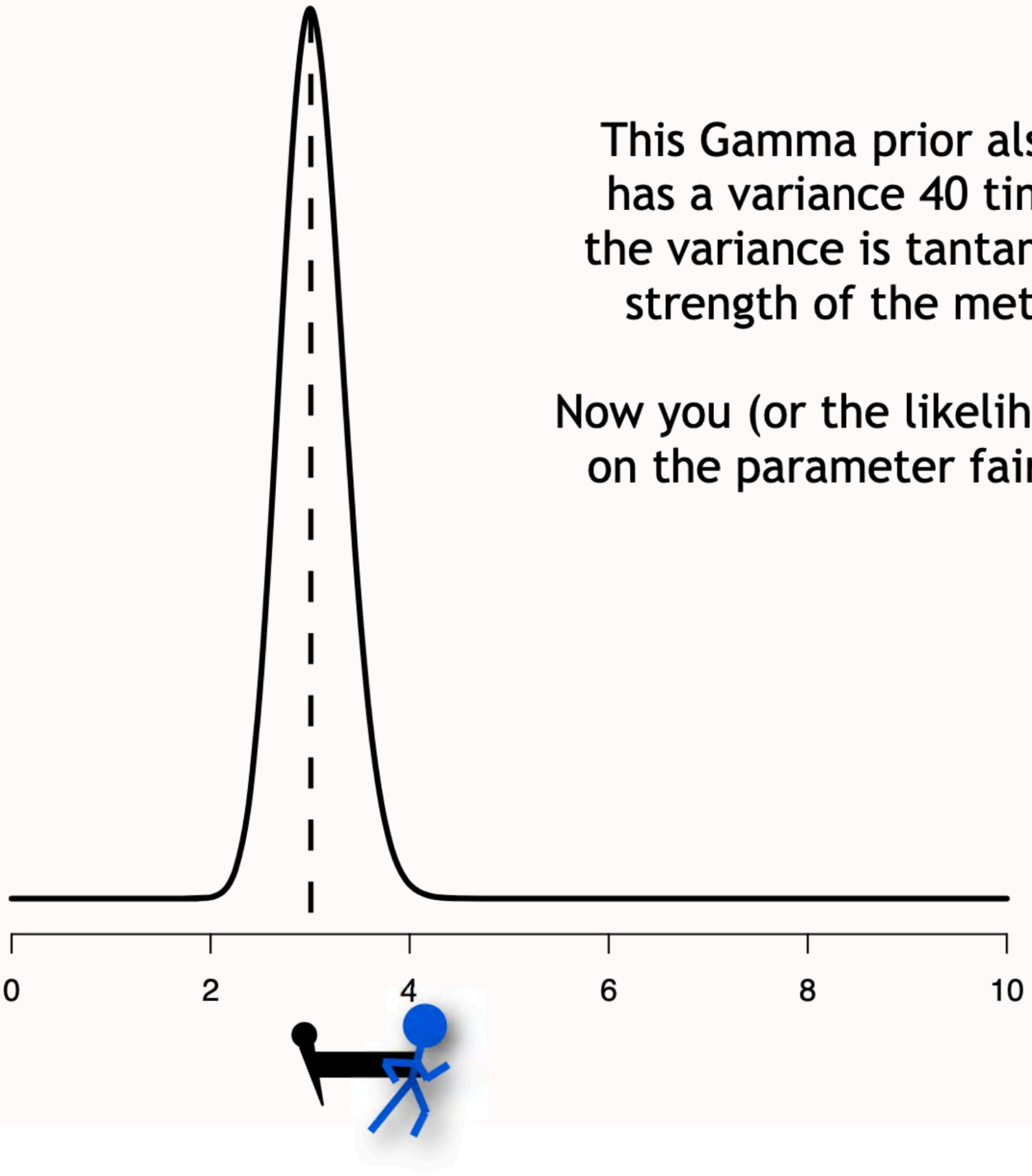
mean = a/(a+b)
variance =
$ab/[(a+b)^2(a+b+1)]$

Non-informative prior



This Gamma(4,1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

The mode of a Gamma(a,b) distribution is (a-1)b (assuming a > 1)

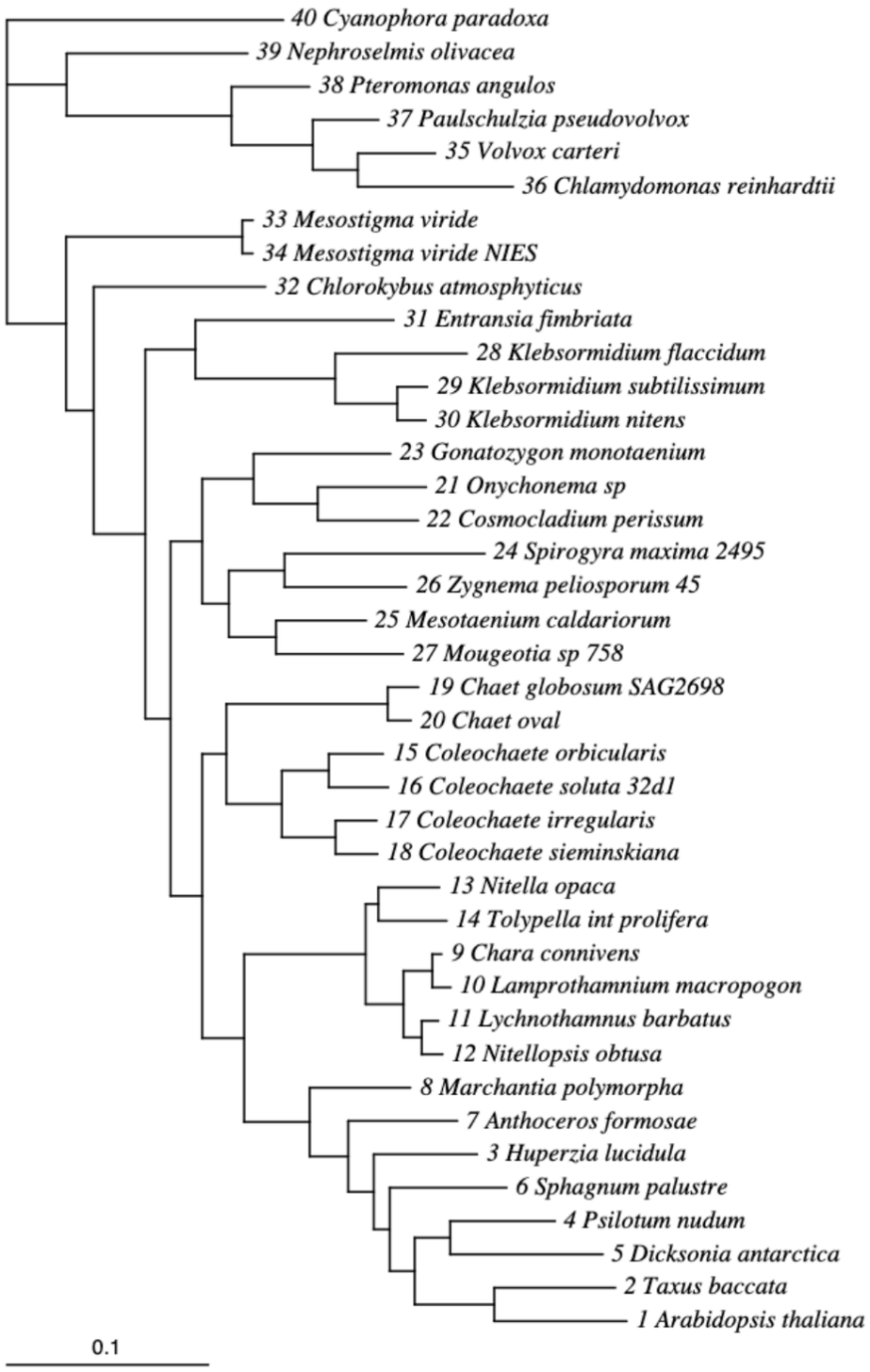


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

Now you (or the likelihood) would have to tug on the parameter fairly hard for it to have a value as large as 4.

This gamma distribution has shape 91.989 and scale 0.032971

Example: Internal Branch Length Priors

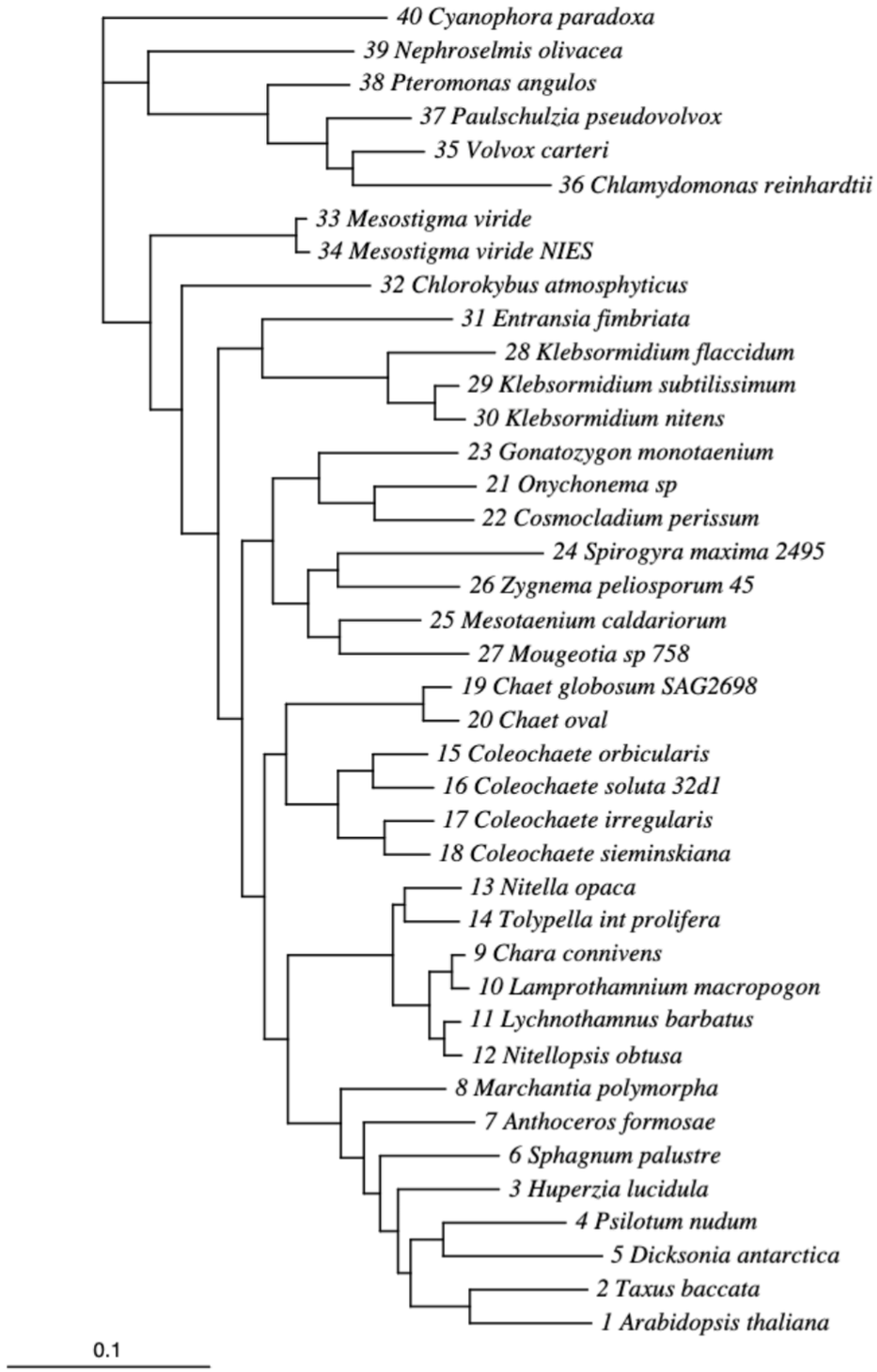


Separate priors applied to internal and external branches

External branch length prior is exponential with mean 0.1

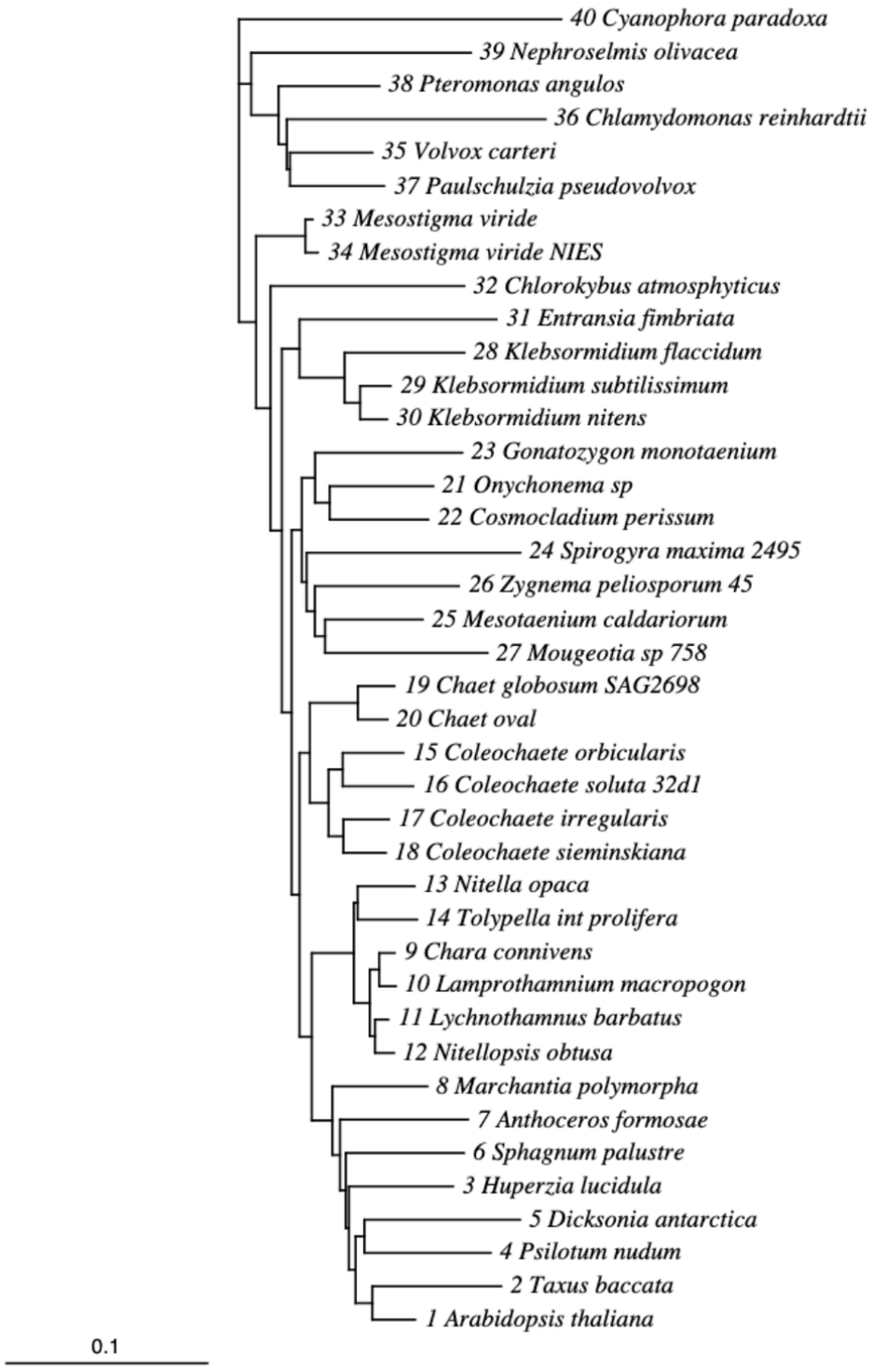
Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior



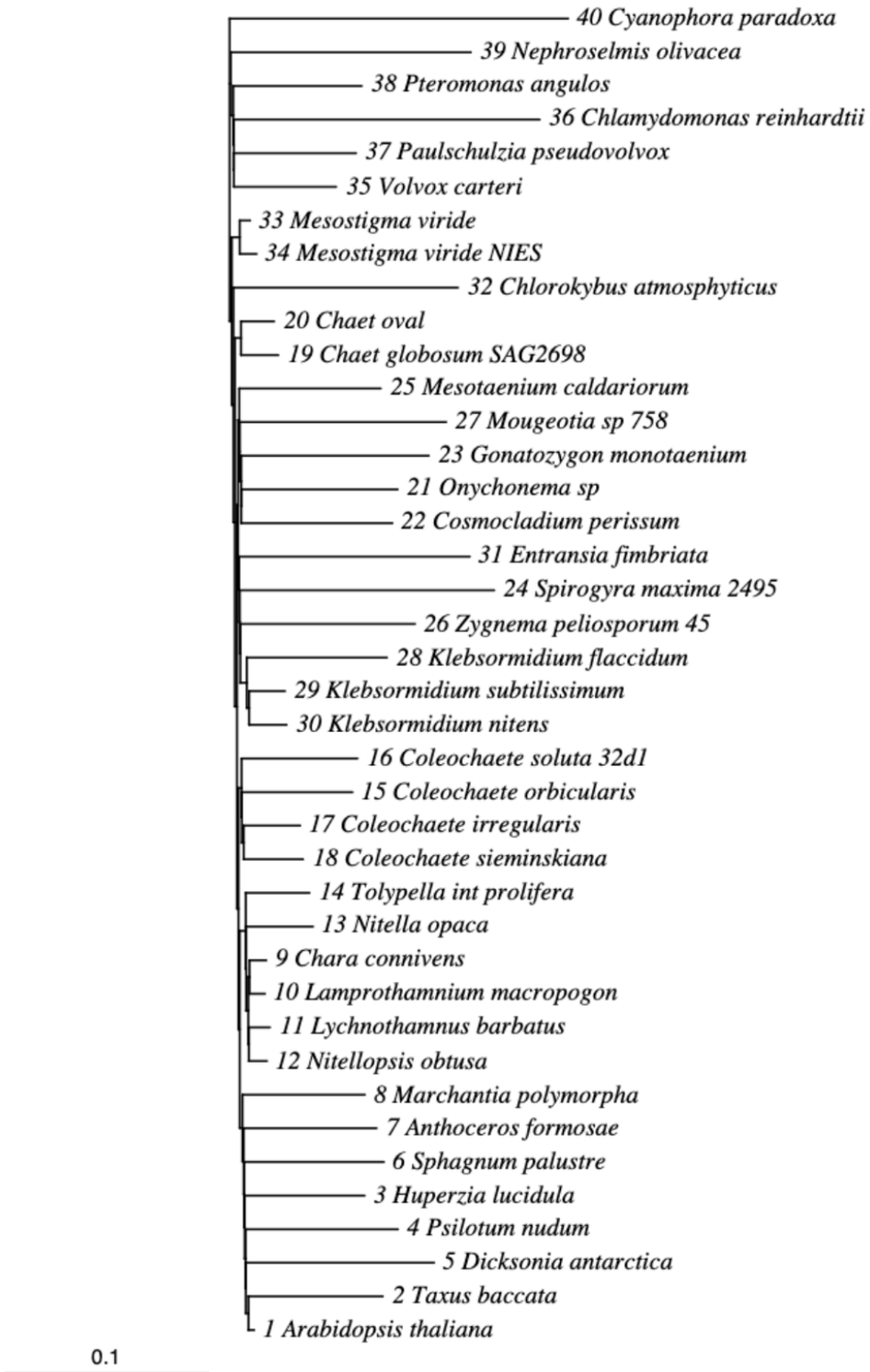
Internal branch length prior mean 0.001

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)



Internal branch length prior mean 0.0001

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)



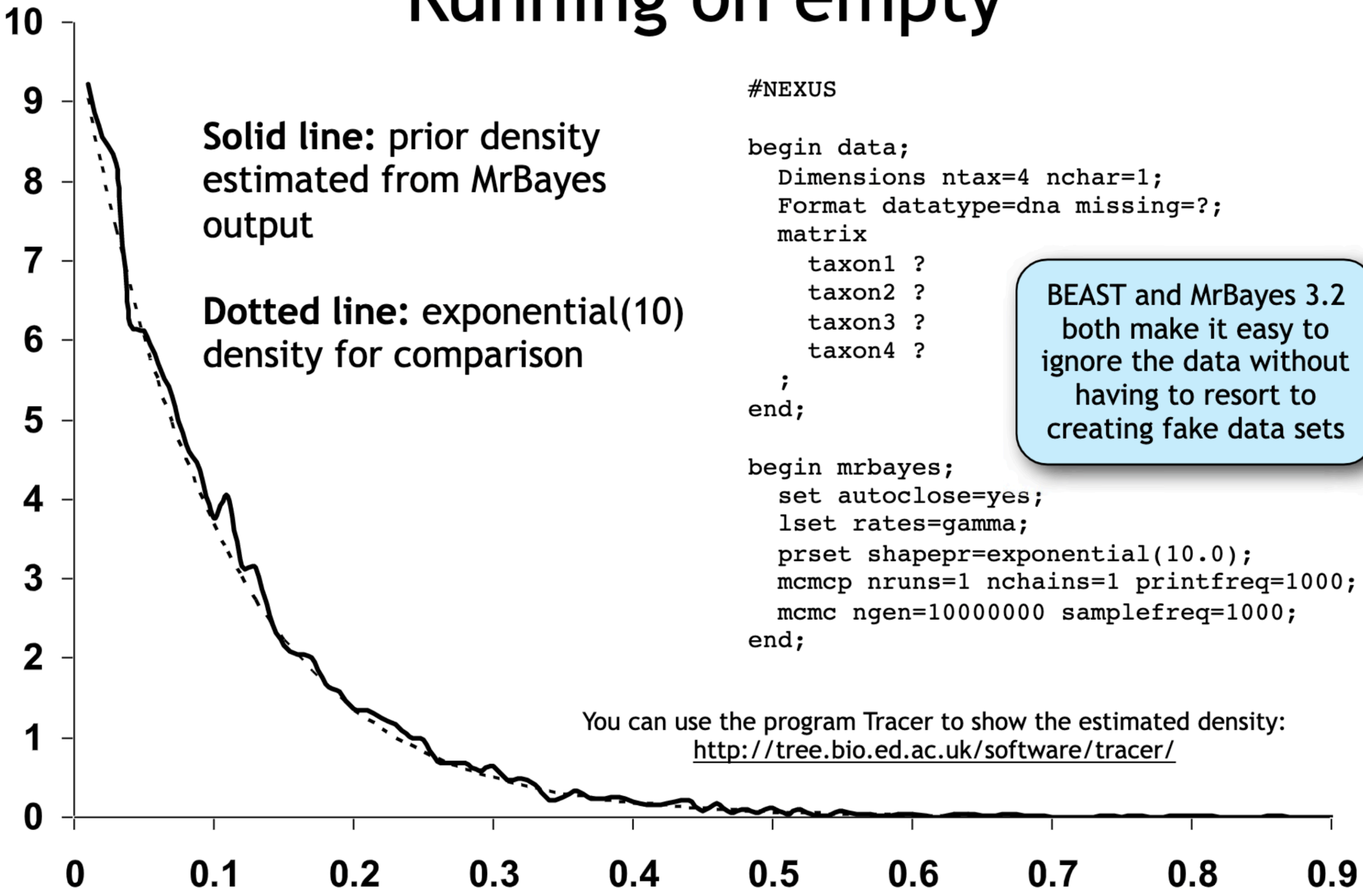
Internal branch length prior mean 0.00001

Running on empty

- Run MCMC without data
- Some programs generate dummy “empty” alignments that can be used
- Why run MCMC on the prior?
    - Check correctness of the software (do the results match the theoretical prior?)
    - Compare prior run with posterior run:
        - If prior and posterior are too similar, then data might contain little information
        - If prior and posterior overlap, but the posterior is more concentrated, then data are informative and prior is reasonable
        - If prior and posterior do not overlap well, then the prior might be misspecified

- CAUTION: do not set a prior to match the posterior! The prior is meant to reflect our knowledge BEFORE the analysis of the data

Running on empty



Solid line: prior density estimated from MrBayes output

Dotted line: exponential(10) density for comparison

```
#NEXUS

begin data;
  Dimensions ntax=4 nchar=1;
  Format datatype=dna missing=?;
  matrix
    taxon1 ?
    taxon2 ?
    taxon3 ?
    taxon4 ?
  ;
end;

begin mrbayes;
  set autoclose=yes;
  lset rates=gamma;
  prset shapepr=exponential(10.0);
  mcmcp nruns=1 nchains=1 printfreq=1000;
  mcmc ngen=10000000 samplefreq=1000;
end;
```

BEAST and MrBayes 3.2 both make it easy to ignore the data without having to resort to creating fake data sets

You can use the program Tracer to show the estimated density:
http://tree.bio.ed.ac.uk/software/tracer/

HW reading

- Yang and Rannala (1997)

For next class:

- We will go over MrBayes
- Each student is assigned to one group:
    - Group 1: Read two papers on MrBayes (short software papers)
    - Group 2: Read Larget and Simon paper
- We will have a class discussion followed by installing and using the software